TRADING BITCOIN WITH REDDIT SENTIMENT: STRATEGIC CAPITAL ALLOCATION USING ML-ENHANCED TECHNICAL SIGNALS

By

Bence Benedek Pál

Submitted to
Central European University
Department of Undergraduate Studies

In partial fulfillment of the requirements for the degree of
BSc Data Science and Society

Supervisor: Márton Pósfai

Vienna, Austria

2025

Copyright Notice

[1]

[1] Icon by Font Awesome: https://fontawesome.com/

Author's Declaration

I, the undersigned, **Bence Benedek Pál**, candidate for the BA degree in Data Science and Society declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 26 May 2025

Bence Benedek Pál

Signature

Abstract

This thesis shows that Reddit crowd mood can sharpen short-term Bitcoin trading when using an appropriate model. The sentiment of three-hundred-thousand finance-subreddit posts from 2017-2024 is analyzed and fused with standard price- and volume-derived indicators. Two empirical tracks test the idea: (i) Random Forest classifiers enhanced with sentiment, (ii) a simplified MACD crossover filtered by machine learning gates that use sentiment features. Both tracks are evaluated on the model-level and the strategy-level. I then statistically quantify the added value of sentiment in Track 1, while also examining how it affects the model's predictions. I find that the sentiment-enhanced Random Forest model outperforms its indicator-only counterpart, while the sentiment-based gates' edge seems to stem only from lowered trading costs. It is also shown that when strong sentiment is paired with high activity levels, we get strong and reliable predictions indicating that Reddit sentiment is a valuable metric in model building. Although the Sharpe-ratio increase becomes statistically insignificant once sampling variance is considered, the evidence still indicates moderate support for sentiment improving forecasting models.

Acknowledgements

I would like to thank my friends and family who have supported me throughout the course of my studies. I would also like to give a special thanks to my supervisor Márton Pósfai, who has helped me complete this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Efficient Market Hypothesis – considered to be a cornerstone of financial theory – suggests that market prices fully reflect available information, making it impossible to consistently achieve abnormal returns and gaining a legal edge over other actors by relying on historical data and publicly available information. However, the modern financial landscape increasingly recognizes the importance of investor market sentiment as a source of market inefficiency. Noise trader theory, developed by De Long et al., argues that sentiment-driven, irrational investors can significantly impact market prices, resulting in deviations from fundamental principles of the market [1]. These traders introduce additional risk into the market, which results in rational arbitrageurs not being able to fully account for such mispricing, resulting in inefficiencies.

The concept of investor sentiment influencing market dynamics has been long theorized by famous economists. Keynes famously introduced the concept of "animal spirits", describing spontaneous waves of investor optimism- and pessimism that guides investment decisions beyond simply rational expectations [2].

This was followed by heavy debates about the efficiency of markets, perhaps until Baker and Wurgler - two pioneers of behavioral finance - demonstrated that sentiment-driven mispricing systematically affect speculative stocks, further supporting that sentiment cannot be discredited as simple noise, but rather as a critical factor in financial markets [3] [4].

Recent technological advancements, especially in data mining and natural language processing (NLP), have strengthened the ability to exploit online sentiment data reaffirming the British mathematician, Clive Humby's famous declaration of data as the "new oil". Such advancements are strongly transformative in a sector like financial markets where incremental edges can result in substantial gains. Platforms, such as Reddit, specifically the r/WallStreetBets subreddit, showcase the importance of investor online activity. The community famously caused unprecedented market turmoil during the GameStop (GME) short squeeze event, where online investor sentiment was shown to be correlated with price movements during the event [5].

Given Reddit's importance as a hub for tech-savvy users and cryptocurrency enthusiasts, it offers a unique insight into investor sentiment towards cryptocurrencies, which are known for

their high sensitivity towards market sentiment and speculation. Bitcoin presents itself to be an ideal asset to investigate sentiment-driven trading strategies due to it being the most prominent cryptocurrency, while still being prone to investor emotions and substantial volatility.

Previous academic research has explored financial sentiment extraction from social media platforms and its predictive potential regarding financial assets, though a never-ending gap will always remain: what is the best way to integrate processed financial sentiment data to construct real-world trading strategies and to build an incremental economic edge. This thesis aims to address this gap by exploring financial sentiment data collected from Reddit discussions to forecast short-term Bitcoin price movements. The subsequent trading strategy utilizes these forecasts to assess whether incorporating sentiment into the models improves predictive accuracy and financial returns, while also investigating how sentiment influences model performance.

The thesis comprises of five - other - main chapters. Firstly, it provides a comprehensive literature review, where it establishes theoretical and empirical foundations of market sentiment analysis and its evolution onto the cryptocurrency markets. Next, it outlines the data collection process and gives a detailed description of the sentiment and market data used. Subsequent methodological sections describe data pre-processing techniques, model selection criteria and architecture details, ensuring transparency and replicability. Following this, model performance is strictly evaluated, and the added value of the processed sentiment is tested through rigorous backtesting. Finally, the thesis concludes by discussing the findings' implications, acknowledging limitations, and situating the research within the broader landscape of behavioral- and quantitative finance literature.

Chapter 2

Literature Review

Sentiment has been studied in different eras of behavioral finance as a potential driver of price movements. In this chapter, I will first introduce research about the prevalence of market sentiment studies to gain an understanding of the fundamentals of behavioral finance literature. Then, I will present the current state-of-the-art AI-driven methods for extracting financial sentiment from text data. Finally, I will extend this notion to the cryptocurrency space, and how it has been used to forecast returns.

2.1 Market Sentiment

Perhaps one of the most important notions of stock market studies is the Efficient Market Hypothesis (EMH), which holds that prices instantaneously reflect all publicly available information leaving no exploitable edge for investors [6]. The psychological foundation for believing that aggregate sentiment can bias prices comes from Kahneman and Tversky's Prospect Theory, which shows that agents overweigh losses relative to gains and rely on reference points, producing systematic, predictable deviations from the perfectly "rational" utility assumed in classical models [7].

The first serious challenge to the EMH was long-run mean-reversion: investors overweigh recent bad news, pushing prices too low for prior losers, while good news pushes prices too high for winners, but when the noise clears, prices correct [8]. Such patterns reveal that markets can misprice assets based on sentiment – something risk-adjusted EMH cannot explain.

This raises the question why the rational actors (arbitrageurs) do not eliminate such mispricing. De Long et al. introduce the idea of noise trader-risk: sentiment-driven traders both cause mispricing and raise the risk for rational actors, so the latter may not fully offset the distortion [1]. This mechanism is especially relevant for volatile, sentiment-sensitive and hard-to-value assets such as Bitcoin (BTC).

Barberis & Thaler analyze large amounts of evidence on biases — overconfidence, loss aversion, representativeness — and link them to persistent anomalies like over- and under-

reaction [9]. They recast sentiment as a systematic driver of mispricing rather than mere noise. Building on that view, the present thesis asks whether and how online sentiment can help predict short-term Bitcoin moves with Bitcoin being a textbook sentiment-sensitive asset.

A key advance in measuring sentiment came when Baker & Wurgler built a monthly sentiment index from market-based proxies of investor enthusiasm and showed that high sentiment predicts lower future returns for young, volatile, hard-to-arbitrage stocks [3]. Although their study uses cross-sectional data and macro scale indicators over a long horizon, it provides the conceptual foundation for this thesis, which examines daily, micro-scale Reddit sentiment.

In a follow-up paper, Baker & Wurgler define investor sentiment as "a belief about future cash flows and investment risks that is not justified by the facts at hand" and argue that its impact is the greatest where assets are difficult to value and to arbitrage [4]. Bitcoin fits this description almost perfectly: no intrinsic cash flows, speculative demand, extreme volatility, non-trivial transfer latency, and fragmented markets. Their framework, therefore, justifies investigating whether sentiment harvested from online activity systematically influences Bitcoin price dynamics — the central question this thesis tackles.

2.2 Text-based Sentiment Analysis in Finance

Tetlock has demonstrated that even a simple negative-dictionary approach to sentiment extraction can be informative: high media pessimism pushes down next-day market prices, while unusually extreme pessimism – high or low – increases trading volume [10]. Extending sentiment mining to social media, Bollen et al. show that Twitter mood, especially the "Calm" dimension Granger-causes the Dow Jones up to a week ahead [11].

The LLM-boom revolutionized sentiment analysis, reaching the financial world with FinBERT, a BERT-based model fine-tuned on finance specific corpus [12]. FinBERT's contextual embeddings capture negation and jargon better than word-count lexicons, yet its equity-news vocabulary misses online, crypto-native slang. To bridge that gap, CryptoBERT – a transformer re-trained on social media cryptocurrency discourse – learns this lexicon and provides three-way polarity scores (bullish / neutral / bearish) [13].

2.3 Sentiment-based Cryptocurrency Trading

As cryptocurrency markets have long been highly speculative because of the absence of fundamentals, they offer excellent opportunities to study whether social media sentiment can translate into the financial market.

Mai et al. show that even a dictionary-based sentiment on Bitcointalk predicts Bitcoin returns, and finds that posts from the silent majority of users carry more weight than the vocal minority [14]. The notion that social media buzz can drive real price movements went main-

stream during the 2021 Gamestop short squeeze, where r/WallStreetBets users coordinated buying to propel the stock. Long et al. confirm that spikes in subreddit activity tracked GME's upward bursts [5].

Most recently, Gurgul, Lessmann and Härdle demonstrate that BERT-classified social media sentiment lifts one-day BTC-return models and boosts the Sharpe-ratio of a simple trade-on-signal strategy – an approach very close to the one adopted here [15]. This thesis tests whether Reddit-derived CryptoBERT sentiment, can improve a five-day Bitcoin trading model beyond what technical indicators achieve alone and whether sentiment alone can serve as a gate on top of an existing technical strategy.

Chapter 3

Data

3.1 Data Sources

The empirical analysis draws on two primary datasets: a social-media corpus of Reddit discussions and daily price data for Bitcoin.

3.1.1 Social Media Corpus

Reddit posts and comments were obtained from the AcademicTorrents [16] public dataset archive, which preserves full JSON data regarding online activity collected via historical API scrapes. Five large, general-finance subreddits – r/WallStreetBets, r/investing, r/stocks, r/options and r/StocksToBuyToday – were selected. These communities outnumber crypto-specific forums in both membership and message flow and, crucially, reflect the broader retail-investor conversation in which Bitcoin appears alongside equities, options, and macro news. Relying solely on niche subreddits such as r/Bitcoin would have yielded a narrower, echo-chamber perspective and a much sparser pre-2020 history.

3.1.2 Market Data

Daily bitcoin prices (ticker BTC-USD) were downloaded from Yahoo Finance [17] using Python's `yfinance` package [18], which reports aggregated 24-hour opening, highest, lowest, and closing prices as well as volume figures. The analysis window is 1 January 2017 – 31 December 2024, as before 2017 Reddit conversational data is too thin for reliable sentiment estimation, while the post-2017 timeframe captures Bitcoin's evolution into an institutional-scale asset.

3.2 Data Cleaning and Sentiment Labeling

All Reddit objects flagged as *[deleted]* or *[removed]* were discarded, together with obvious Discord-invite spam, as none of them carry substantive content. The remaining dataset was fil-

tered with a light keyword screen - `"bitcoin"`, `"btc"`, `"binance"`, `"crypto"` (case-insensitive) - to retain threads in which Bitcoin is explicitly discussed. Every comment nested under a qualifying submission was then included, even if the comment itself lacked the keywords, to preserve the conversational context and to avoid skewing sentiment towards the opening post. After the filtering step, the dataset comprises of 11 090 posts and 274 162 comments.

Message-level sentiment is extracted with CryptoBERT, a BERTweet-based model fine-tuned on crypto-centric social media text and emoji polarity lexicon [13]. The model is able to handle multilingual slang, typical in these forums. Messages longer than the model's 512-token window ($\approx$ 2000 characters) are truncated.

CryptoBERT returns softmax probabilities $P_{pos}, P_{neu}, P_{neg}$ for the three tone classes. Each message is assigned a label corresponding to $max(P)$ and a polarity score:

$$polarity = P_{pos} - P_{neg},$$

which ranges from -1 (strongly bearish) to +1 (strongly bullish). Finally, scores are aggregated on the daily level, producing sentiment features:

- mean daily polarity,
- total message count,
- positive and negative count ratios

3.3 Features and final dataset

3.3.1 Market-side engineering

From the BTC-USD series, I derive a compact but informative feature set. For each rolling window $w \in \{10, 15, 20\}$ I compute:

- z-standardized closing price $Z_w(\text{Close})$ and trading volume $Z_w(\text{Vol})$,
- the w-day simple moving average of both variables,
- w-day volatility
- $\text{RSI}_w - 50$[1] (centered to zero for comparability),
- momentum M_w and its five-day difference $\Delta_5 M_w$,
- five-day percentage change in closing price $\Delta_5\text{Close}$

[1]The Relative Strength Index (RSI) is a bounded (0-100) oscillator that compares average gains to average losses over a look-back period, with readings above 70 typically labeled "overbought" and below 30 "oversold"

- five-day differences of volume and RSI, and
- the intraday return: closing price minus opening price

3.3.2 Sentiment-side engineering

For each day t, using the computed polarity score p_t^{daily} and the window $w \in \{10, 15, 20\}$, I compute:

- daily mean polarity p_t^{daily},
- message count n_t (”activity”),
- $Z_w(p_t^{\text{daily}})$ and $Z_w(n_t)$,
- $\log\left(\frac{\text{pos ratio}}{\text{neg ratio}}\right)$,
- a ”max-ratio” feature: the larger of the positive or negative share (assigned a negative sign if bearish),
- five-day differences $\Delta_5 p_t^{\text{daily}}$ and $\Delta_5 n_t$,
- two lags of polarity: $p_{t-1}^{\text{daily}}, p_{t-2}^{\text{daily}}$, and
- an interaction term $p_t^{\text{daily}} \times n_t$.

Note that Ider & Lessmann employ a normalized sentiment score defined as $(\text{pos} - \text{neg})/(\text{pos} + \text{neu} + \text{neg})$ [19]. In contrast, this thesis retains the simpler probability difference $P_{\text{pos}} - P_{\text{neg}}$, which is bounded, interpretable, and requires no additional transformation.

3.3.3 Merging and Sample

Because both Yahoo Finance and Reddit provide UTC based data, features are merged on the calendar date without further alignment. After dropping rows with any missing feature, the final dataset contains 2 551 daily observations:

- Training: 1 827 days (2017-01-01 $\rightarrow$ 2022-12-31; 345 sentiment-sparse days naturally fall out)
- Validation: 365 calendar days (year 2023)
- Test: 360 days (year 2024)

The following train/validation/test split follows the walk-forward structure used in recent crypto-forecasting and follows Bitcoin's evolution to a mainstream asset [19] [15]. On Figure 3.1 we observe that the training period captures multiple boom–bust cycles, the validation year covers the 2023 consolidation, and the test window coincides with the 2024 ETF-driven hype.

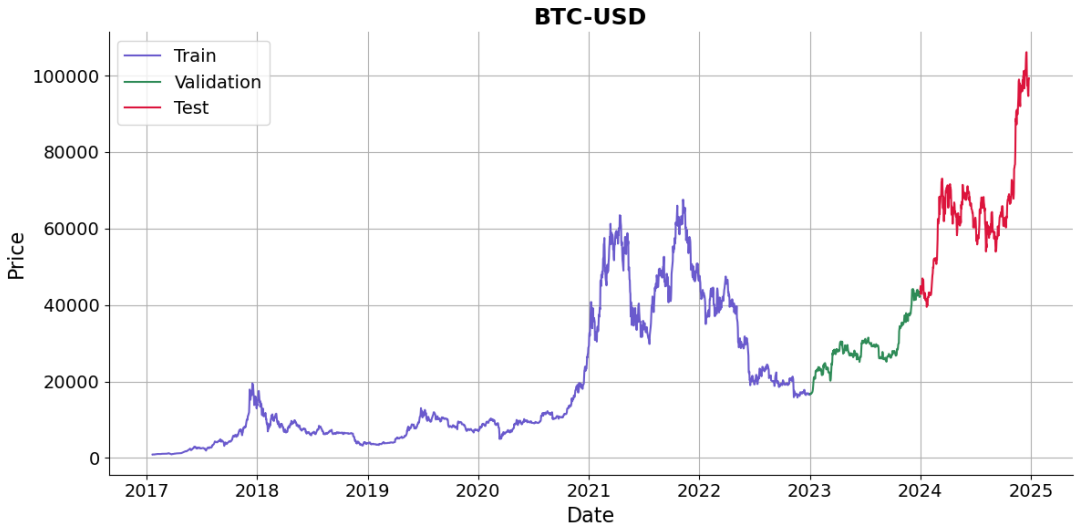

Figure 3.1: BTC-USD with shaded regions according to the split

3.4 Descriptive Statistics

The top panel of Figure 3.2 shows that daily polarity is highly volatile in the early years — reflecting sparse message counts — yet clusters around a moderately bullish range after 2020, with only a handful of extreme spikes. The series is overall shifted upward, indicating a net optimistic tone across the full sample.

The lower panel of the same figure documents message activity. Volume is negligible until late-2017, then oscillates between bursts and quiet periods; heteroskedasticity of the variable justifies the inclusion of activity and interaction terms in the forecasting models.

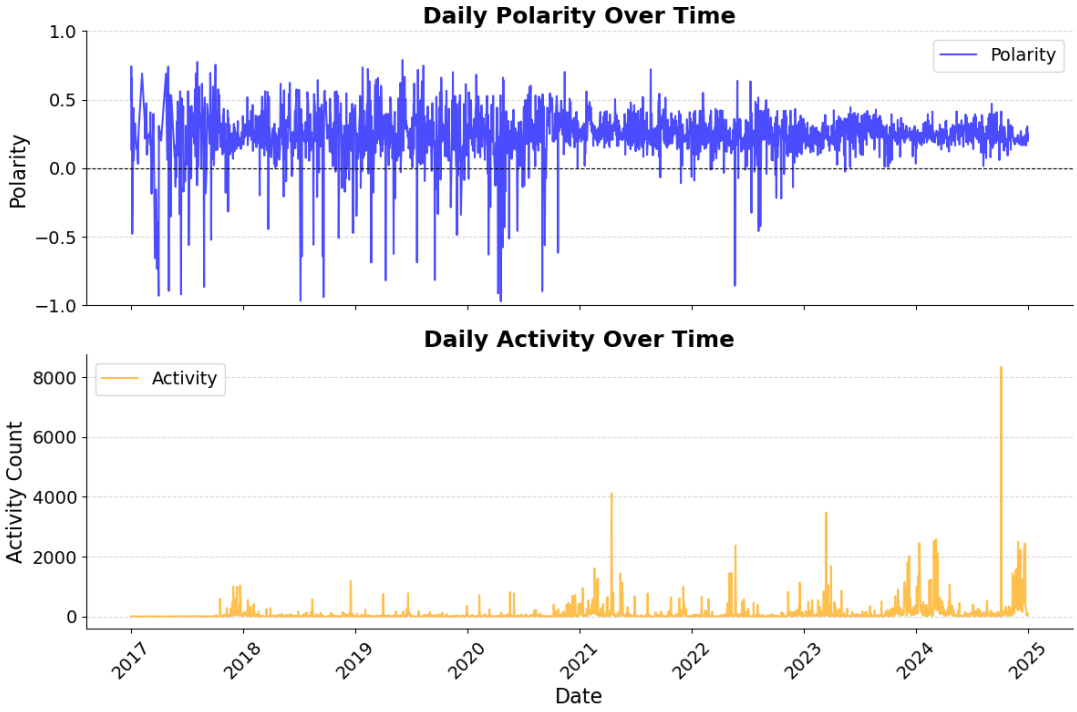

Figure 3.2: Daily polarity (top) and Reddit activity (bottom), 2017–2024.

Yearly aggregates in Figure 3.3 underline the same point: message volume more than doubles from 2020 to 2021 and again from 2023 to 2024, mirroring Bitcoin's shift from a niche asset to mainstream topic.

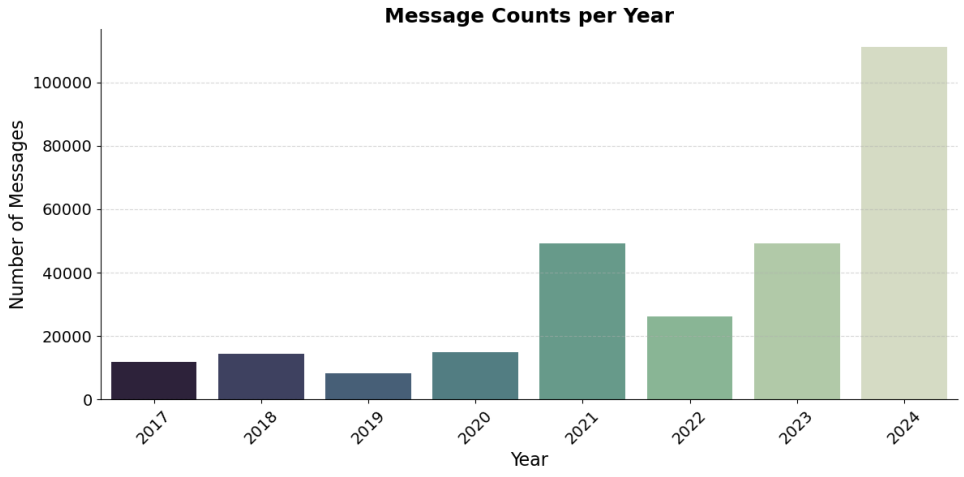

Figure 3.3: Total Bitcoin-related messages per calendar year.

To understand the nature of the sentiment polarity scores used in this analysis, we can look at Figure 3.4, which plots the unconditional distribution of daily polarity. A strong left tail,driven by low-volume days in 2017-18 creates a negative skew, yet 95 % of observations remain positive, consistent with the generally positive, hype-generating sentiment around Bitcoin on social media.

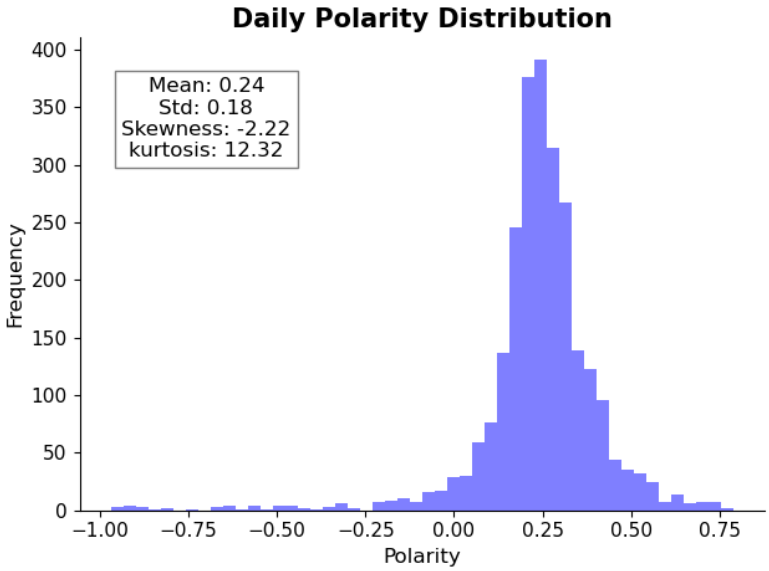

Figure 3.4: Histogram of daily polarity ($n = 2551$).

Chapter 4

Methodology

4.1 Research Design Overview

The empirical study proceeds on two parallel yet complementary tracks. Track 1 treats return prediction as a pure machine learning exercise. Six Random Forest (RF) classifiers are trained, each fed with indicators calculated using the different window lengths. For every horizon, two RF variants are built: one that considers the full feature set, and a benchmark that excludes sentiment, thus isolating the value of investors' mood. All models follow a walk-forward protocol: they are trained on data from 2017-2022, thresholds are calibrated on 2023, and out-of-sample performance is assessed on 2024. The effect of sentiment is analyzed with regards to the results and statistical significance is quantified.

Track 2 begins with a shortened Moving Average Convergence Divergence (MACD) crossover rule – defined as the point at which the difference between a short- and a long-term exponential moving average (EMA) of the closing price ("MACD line") crosses its five-period EMA ("Signal line"), a configuration considered to be a very basic algorithmic trading strategy. Each raw long signal with a one week holding period is treated as a candidate trade resulting in 162 signals over the sample. Because this sample is too small for a conventional train-test split, the gate classifiers are evaluated with leave-one-out cross-validation (LOOCV): for each crossover the model is trained on all remaining signals and then asked to predict the left-out observation. Three filters are benchmarked against the unfiltered MACD: Gaussian Naïve Bayes, ridge-penalized logistic regression, and a depth-3 Random Forest, yielding 4 MACD variants in total. The LOOCV procedure supplies out-of-sample predictions for every signal, which are then fed directly into the back-test. The results of this track are not quantified statistically and its sentiment effect is not discussed in-depth as Track 1 offers data depth, feature richness and strong performance that make inference and interpretability worthwhile, whereas Track 2 does not.

A uniform transaction-fee assumption of 0.075% applies to every buy and sell, reflecting the

commission paid by a typical Binance user who settles fees in Binance Coin (BNB).[1] Once the models are locked, the resulting predictions are assessed in two layers: first at the model level then subsequently at the strategy level, so that statistical accuracy and economic usefulness can be distinguished.

4.2 Feature Engineering

Financial forecasting research consistently shows value in combining standardized technical indicators with sentiment features. Studies support using a mix of standardized technical indicators and sentiment features. Arian et al., for example, use z-scored price series to isolate price anomalies and augment them with momentum signals such as RSI and rate-of-change [20]. Fixing such features ex-ante curbs data mining: as Bailey et al. note, the probability of back-test overfitting rises with every additional rule or parameter tried [21]. Parallel work on investor mood reports similar gains. Mai et al. show that social-media sentiment significantly predicts Bitcoin returns, while Chen et al. and Nasekin & Chen find that sentiment scores sourced from Reddit or StockTwits improve crypto-return forecasts when extracted in a finance-specific context [22] [14] [23]. These findings justify including sentiment metrics – polarity, activity, and their derivatives – alongside traditional price- and volume-based indicators. Locking the feature set before modeling thus mitigates multiple testing bias - as emphasized by Bailey et al. - and aligns with current academic practice in cryptocurrency forecasting, helping to ensure that any predictive edge stems from economic underpinnings rather than from fitting to noise in the sample [21].

4.3 Random Forest Model

Tree-based ensembles have been widely used in return prediction due to their proven performance in noisy financial domains and their robustness: Krauss et al. show on U.S. equities that a Random Forest model outperforms single learner models like logistic regression, boosted trees or deep networks in terms of risk-adjusted returns [24]. Random Forests handle high-dimensional, non-linear data well and are relatively resistant to overfitting, which makes them suitable for the high-noise, non-linear patterns of the cryptocurrency markets. Similarly, Kalyani et al. found that including news sentiment increased prediction accuracy, with a Random Forest classifier achieving the highest accuracy among the models tested [25].

As class imbalance is common and it can result in biased models, I employ Python's `scikit-learn`'s built-in class balancing method. I use walk-forward time series cross validation as traditional k-fold cross validation can leak information in time-series, leading to

[1] This rate corresponds to the spot commission of a regular user of 0.10%, reduced by 25% discount applied when fees are paid in Binance coin (BNB)

overly optimistic estimates. Hyperparameter optimization of the models was done on the training set to preserve the validation set for threshold tuning. By balancing classes and employing time-aware cross-validation, my Random Forest classifiers follow best practices recommended in the literature for dealing with imbalanced financial market data.

4.4 Calibration and Threshold Selection

For the machine-learning track, I first calibrate the model's predicted probabilities and then select a trading threshold optimized for economic performance. To calibrate the output of the hyperparameter tuned model, I apply Platt-scaling via a logistic sigmoid using a `CalibratedClassifierCV` in `prefit` mode, which fits the sigmoid on the training data without refitting the classifier itself. This is important because the raw class-vote fractions from the RF model typically cluster around 0.50 in noisy classification problems, not reflecting an interpretable, true confidence of predictions. Calibration stretches these compressed scores onto a meaningful probability scale, so that the predicted probability reflects the likelihood of a correct classification [26].

After calibration, I drop the default 50% cut-off and search for a probability threshold that maximizes the Sharpe-ratio on the validation year 2023. The Sharpe-ratio[2] is computed using a constant 5% annual risk-free rate, chosen to approximate the yield of the 3-month U.S. Treasury bill across 2023 and 2024 – an appropriate funding benchmark for a USD-based trader in that period.

The threshold search is constrained to the interval $[0.50, 0.70]$: values below 0.50 would authorize trades whose implied edge is negative, while higher cut-offs risk overfitting to the validation set. Tuning the decision threshold on a reserved validation slice - using Sharpe-ratio as the score to beat - lets the model optimize directly for the risk-adjusted return, on which it will later be judged upon, while guarding against look-ahead bias. The procedure also reflects Fawcett's point, namely that a classifier's decision rule should align with the application's payoff structure [27].

4.5 MACD Signals and Gating Classifier

As outlined in Section 4.1, Track 2 starts from a short-horizon MACD crossover and then passes each raw long signal through a machine-learning "gate".

[2]The Sharpe-ratio measures risk-adjusted performance, it equals a portfolio's average return minus the risk-free rate, divided by the standard deviation of that excess return series

4.5.1 Gating a Signal

MACD remains one of the most cited momentum rules in practice, and early evidence showed that simple moving-average systems earned higher than expected returns on U.S. equities [28]. Later work - however - found that once realistic trading frictions are introduced the edge often vanishes [29]. At the same time, market-specific tuning can restore significance: a hyperparameter-tuned MACD produced positive alphas[3] on the Nikkei 225[4] [30]. These mixed findings motivate a hybrid approach: keep the fundamental value encoded in MACD, but let a data-driven gate decide when the signal is credible.

4.5.2 Signal Construction

In place of the textbook 12-/26-/9 configuration, I use a $8-/17-5$ MACD: the MACD line is spread between an 8-period and a 17-period EMA, while its five-period EMA serves as the signal line. The shorter windows accelerate the indicator's reaction to price changes and - crucially - yield more (162 in this case) bullish crossovers between 2017 and 2024 - enough observations to train and validate the gate with leave-one-out cross-validation, yet still sparse enough to avoid over-trading. Each crossover triggers a candidate long position held for exactly five trading days, the gate then decides whether the trade is executed.

4.5.3 Gate architecture

Each MACD crossover is analyzed by the three classifiers, run in parallel as separate models:

- a Gaussian Naive Bayes
- a ridge-penalized logistic regression, and
- a depth-3 Random Forest.

To guard against over-fitting on the small sample of roughly 160 signals, all gates use the same deliberately slim feature set, created solely from online sentiment introduced in 3.3.2:

- $Z_w(p_t^{\text{daily}})$: the w-day z-score of Reddit polarity,
- $\Delta_5 p_t^{\text{daily}}$: the 5-day change in polarity
- $Z_w(n_t)$: the w-day z-score of posting activity,
- $\%\Delta_5 n_t$: the 5-day pct change in activity, and
- $p_{t-1}^{\text{daily}}, p_{t-2}^{\text{daily}}$: two lags of polarity.

[3] Return that exceeds what would be expected from standard market-risk factors

[4] The Nikkei 225 is the principal equity index of the Tokyo Stock Exchange and a common benchmark for Japanese equity performance

No price-based inputs are included, ensuring that the gate focuses on sentiment.

The Gaussian Naive Bayes and ridge logit serve as transparent baselines. The Random Forest is added for nonlinear flexibility but is heavily regularized:

- `max_depth = 3`, and `min_samples_leaf = 16` ($\approx 10\%$ of the data) curb tree complexity,
- `n_estimators = 200` provides a stable ensemble without excessive variance, and
- `class_weight = "balanced"` offsets the modest class imbalance

4.6 Back-testing Framework

The back-testing setup follows careful standards outlined in academic literature to ensure a realistic and time-consistent evaluation of trading strategies. I implement a walk-forward (rolling) validation framework, meaning the model is retrained and tested sequentially over time. This structure guarantees that all training data strictly precede the validation and test windows, avoiding any look-ahead bias [31]. To combat this from the feature side, every buy signal is also shifted by one period, to rule out look-ahead, when evaluating the portfolio performance.

A new position is not initiated if an existing trade is still active to account for the overlapping trade constraints. This prevents double-counting of returns and enforces a simple budged constraint, which would otherwise be violated by stacking similar positions. While it is theoretically possible to implement dynamic position sizing – adjusting exposure based on model confidence – this introduces complexity beyond the scope of the current thesis.

Each trade is held for a fixed 5-day period in the first track and for one week in the second track, a standard simplification that helps to isolate the effect of entry signals without complex exit rules [20]. Although a fixed horizon may not always align with the optimal exit in hindsight, it makes the simulation framework consistent and allows for a cleaner analysis for the predictive value of sentiment.

Additionally, I apply realistic transaction costs in all simulations to avoid overstating profitability. Including transaction costs prevents overly optimistic Sharpe-ratios and ensures that all reported performance metrics are viable under real-world trading scenarios.

Finally, model performance is evaluated exclusively on the out-of-sample test set, to ensure that reported results are unbiased and statistically honest, in line with best practices in algorithmic finance research.

4.7 Evaluation and Statistical Inference

The evaluation of the models and trading strategies follows a multi-layered approach, to assess both the predictive performance of the models and the economic usefulness while accounting

for uncertainty in both.

I begin by inspecting the standard classification metrics of the models, including precision, recall, and f1 scores. While these metrics are informative, it is important to note that the models were not tuned to optimize them – thresholds were selected to maximize the Sharpe-ratio on the validation set. As such, these classification scores serve as diagnostic tools rather than evaluation criteria.

To understand which variables drive model predictions, I analyze the feature importance of the trained classifiers. Special attention is given to sentiment-based features, in order to assess their role relative to traditional indicators. To understand how they influence the models' predictions, I examine SHAP values[5] for some of the features [32]. SHAP values help explain how much each feature - including sentiment - influences a model's prediction, showing whether, when and how sentiment plays a meaningful role in the model's decision.

Next, I evaluate each model's trading performance through back-testing. The resulting strategies are compared using a standard set of performance metrics: Sharpe-ratio (annualized), maximum draw-down[6], hit-rate (proportion of profitable trades), number of trades, total return and mean net return per trade. These metrics provide a broad view of the quality of the strategy and help to understand the added value of sentiment.

To assess the statistical significance of the performance difference in the first, machine learning track, I employ an IID bootstrap-based inference procedure [33]. Specifically, I resample trade-level returns for the sentiment-augmented and indicator-only strategies, compute their Sharpe-ratios in each resample and build a bootstrap distribution of the Sharpe-ratio differences across 10 000 simulations. This follows the spirit of Brock et al., who use bootstrap simulations to analyze the statistical significance of trading-rule returns, but I extend this idea to the Sharpe-ratio inference [28]. This lets me construct confidence intervals for the difference and quantify the range of plausible outcomes under the observed data-generating process.

Furthermore, I use this sampling distribution in a Bayesian-like inference framework: by computing the proportion of replicates where the sentiment model outperforms the baseline, I estimate the probability that the sentiment-enhanced strategy is superior. This approach does not rely on asymptotic normality and helps correct for data-driven model selection and parameter uncertainty, two common sources of bias in financial forecasting.

Together, these evaluation techniques aim to provide a robust and interpretable assessment of model performance – both from a predictive standpoint and in terms of economic-value. The goal is not only to compare strategies, but to understand which components drive performance and how confident we can be in those findings.

[5]SHAP (Shapley Additive Explanations) values decompose a model prediction into additive contributions from each feature, reflecting their marginal effect under all possible feature permutations.

[6]Maximum draw-down measures the largest peak-to-trough decline in portfolio value during the back-test, capturing the worst observed loss from a previous high.

Chapter 5

Results

5.1 Model-Level Predictive Performance

5.1.1 Machine Learning Track (Track 1)

After the Sharpe-ratio optimized cut-off is applied, each classifier emits a binary 5-day buy signal. Figure 5.1 plots the resulting error trade-off: the horizontal axis showcases losing trades and the vertical axis displays the missed opportunities. For the 10-day look-back the sentiment-enhanced model cuts false-positive-rate (FPR) by roughly one-third, while keeping the false-negative-rate (FNR) essentially flat, a shift that reduces costly losers without surrendering many winners.

At 15 days, the indicator-only model becomes so conservative that it fires almost no trades yet misses nearly everything that matters. Including the sentiment features relaxes the setup just enough to lower the false-negative rate by about twenty percentage points at the cost of a slight increase in the false positives, a trade-off that increases the exposure to profitable periods. At the longest window – 20 days – sentiment dominates outright: lower FNR and lower FPR.

Figure 5.2 converts those error profiles into precision, recall, and F_1 metrics. Precision is higher for the sentiment model at every horizon and increases with widening of the windows,

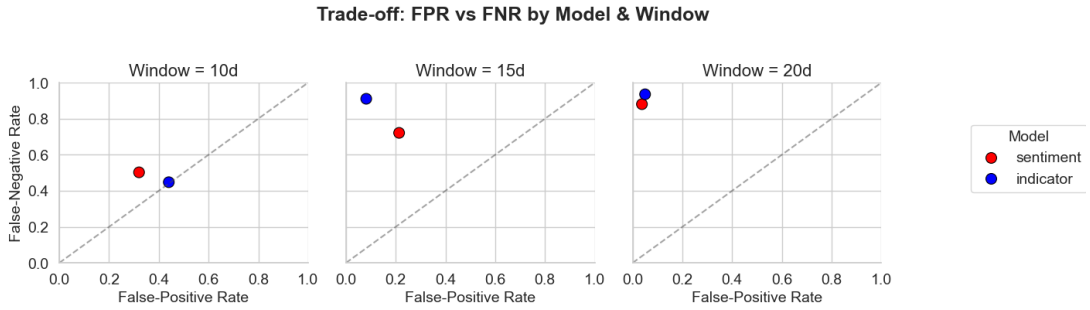

Figure 5.1: Tradeoff between the false-positive-rate (FPR) and the false-negative-rate (FNR) for the different models at the different windows

confirming that the trades it executes are rarely "duds". Recall, by contrast, collapses for the purely technical model as the look-back lengthens, whereas the sentiment-enhanced model moderates the fall (even increasing recall at 15-days to 20-days). The combined effect is a marginally lower F_1 score at the 10-day window, but a decisive advantage once the window widens.

Because the thresholds were tuned for Sharpe ratio, both classifiers favor precision over trade frequency. The sentiment-enhanced version, however, chooses its additional positions carefully, recapturing a part of the upside that the indicator rules leave on the table without flooding the strategy with noise. This selective boost becomes more valuable as longer windows dilute the predictive power of price- and volume-derived indicators alone. The cleaner error characteristics and higher F_1 scores therefore foreshadow the moderate but economically meaningful lift in risk-adjusted returns, quantified in Section 5.2.1.

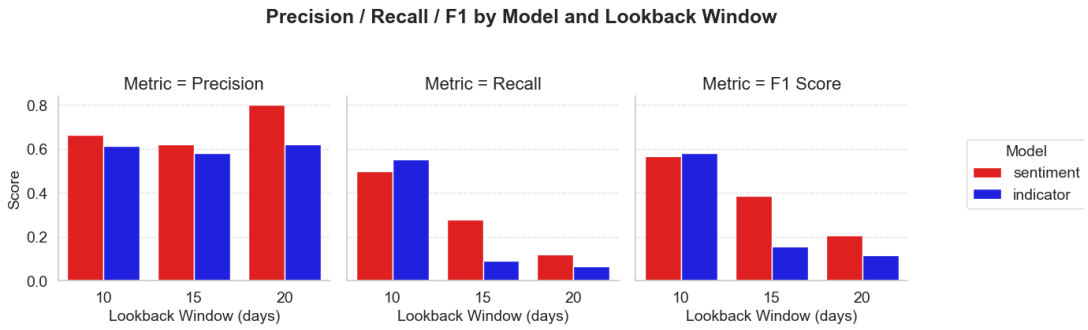

Figure 5.2: Precision, Recall and F_1 for the models with the different look-back windows

5.1.2 MACD and Gate Performance

As outlined earlier, (Section 4.1) the gating classifier is trained only on observations that coincide with a MACD crossing. With no filter in place, the simple MACD therefore labels every instance as positive, achieving a recall of 1.0, but a precision of roughly 0.45 [Figure 5.3] – the latter simply reflects the base-rate of profitable signals in the sample (Figure A.3).

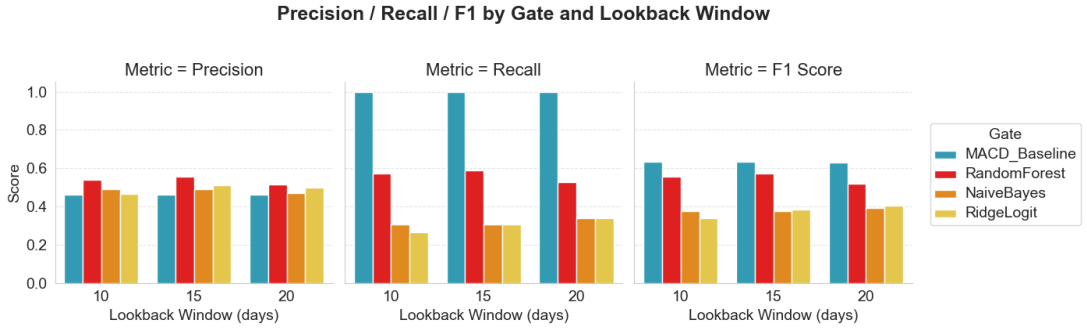

Figure 5.3: Precision, Recall and F1 of the models with the different look-back windows

Ridge-Logit

The ridge-regularized logistic model acts as a very conservative gate: it blocks the vast majority of MACD triggers – good and bad alike – and ends up discarding roughly 80% of the trades that would have made money [Figure 5.4]. This pattern suggests that profitable and unprofitable MACD events are not linearly separable in our sentiment feature space (see Section 3.3.2).

Random Forest

In contrast, thanks to its non-linear splits, the forest is consistently the best performer among the sentiment-only gates. Across 10-, 15-, and 20-day look-back windows it sustains a precision of about ≈ 0.55, recall of ≈ 0.60, and an F_1 of ≈ 0.55. This shows, that a modest smoothing of the sentiment inputs does not erode the ensemble's edge, meaning that the model appears to capture both higher-frequency and regime-level shifts equally well.

Naive Bayes

The Bayesian classifier performs much closer to the ridge-logit rather than to the random forest. Precision stabilizes in the mid-40% range, but recall never rises above 0.35, thus F_1 remains around 0.40 regardless of window length. Evidently, the strong conditional-independence assumption leaves too much signal unexplained, and additional smoothing neither helps nor hurts.

Implications

Sentiment alone presents itself as no silver bullet, but a flexible non-linear gate that can prune many losing MACD signals without sacrificing many winners. The forest roughly halves the trade count, while nudging the hit rate upward - an efficiency gain that could translate into lower transaction costs. Whether these modeling advantages persist once fees and slippage are taken into account is tested in the back-test reported in Section 5.2.2.

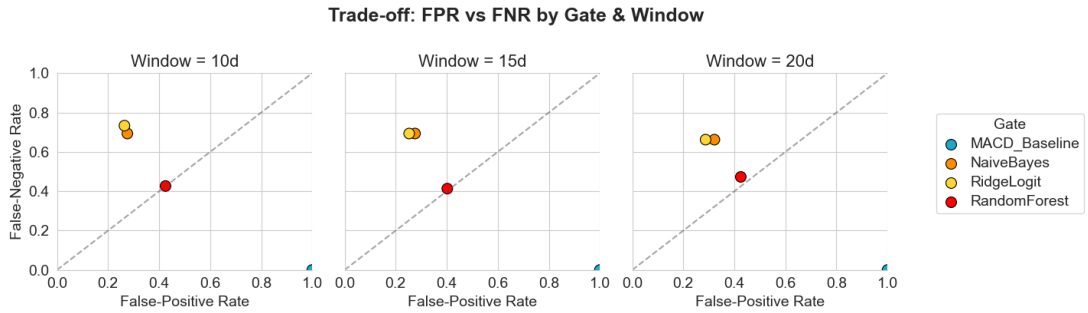

Figure 5.4: Tradeoff between the false positive rate (FPR) and the false negative rate (FNR) for the different models at the different windows

5.2 Strategy Back-test Metrics

5.2.1 RF Indicator-only vs Sentiment-enhanced Back-test

On the strategy level, I have back-tested the predictions of the model. Figure 5.5 plots the equity curves with the starting capital of 100.000 USD for all six random forest strategies – three look-back windows, with and without sentiment features – alongside a buy-and-hold benchmark, that showcases the performance if someone bought at the start of the period and sold at the end of the period. Table 5.1 summarizes trade count, hit rate, mean net return per trade, annualized Sharpe ratio, maximum draw-down and total return. I must note that the performance of the buy-and-hold leads to well-performing strategies as the period is upward trending.

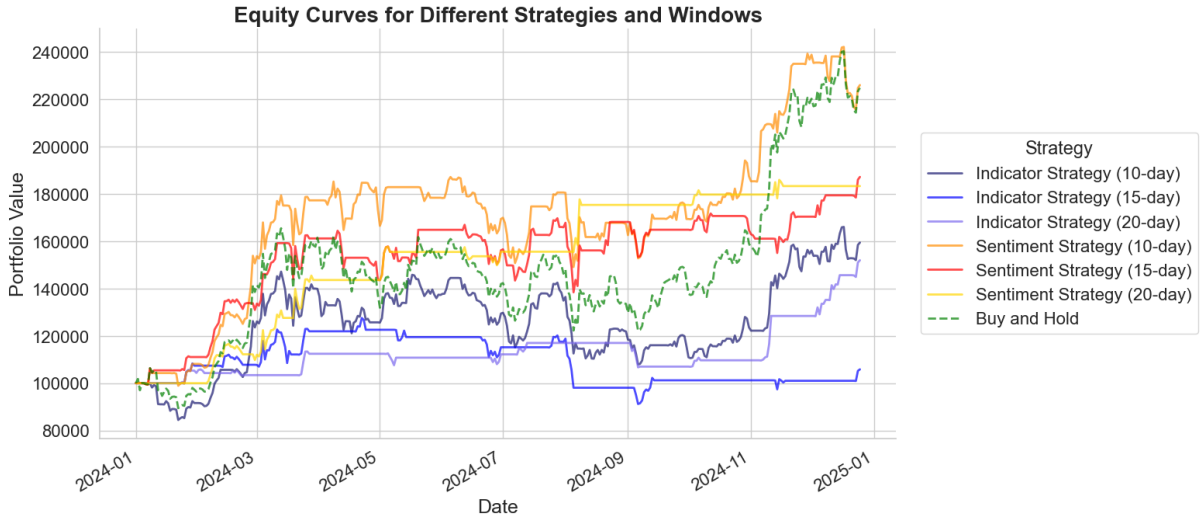

Figure 5.5: Simulated Portfolios for the different models with different window sizes

Table 5.1: Performance Metrics by Window and Model Type

Window	Model-Type	Trades	HitRate	MeanNet	Sharpe	MaxDD	TotalReturn
10d	Indicator	57	0.5614	0.0111	1.1369	-0.2664	0.5944
10d	Indicator+Sentiment	48	0.6875	0.0207	1.9882	-0.2033	1.2602
15d	Indicator	17	0.5882	0.0042	0.1596	-0.2850	0.0585
15d	Indicator+Sentiment	34	0.5588	0.0208	1.7443	-0.1845	0.8719
20d	Indicator	13	0.6154	0.0317	2.0572	-0.0880	0.5197
20d	Indicator+Sentiment	14	0.8571	0.0468	2.3022	-0.0663	0.8332
–	Buy & Hold	–	–	–	1.7083	-0.2618	1.2482

10-Day Window

The sentiment enhanced strategy essentially screens out the bad trades, which doubles the total return and mean trade profit. Sharpe rises from 1.14 to 1.99, while the Maximum Draw-down

improves by about 6 percentage points (pp), while also reducing the number of trades. Fewer, cleaner trades free up capital for other uses – an attractive property in a multi-asset setting.

15-Day Window

Both RF-models stumble, but the indicator-only model collapses, leading to a Sharpe of 0.16. The sentiment model still delivers a respectable 1.74 Sharpe and a 6 pp lower draw-down, turning capital into an 87% gain. This indicates that a 15-day look-back window is long enough to blur short-term swings yet too short to capture regime shifts. Sentiment partially compensates but cannot recover the lost edge entirely.

20-Day Window

Here the sentiment model is the stand-out winner. With only 14 trades it yields an 83% return an 86% hit rate, and the best risk profile (Sharpe = 2.30, MaxDD = -6.6%). The longer window seems to align with macro mood cycles and the non-linear nature of the RF can spot large-move opportunities and sit out volatile markets. The indicator-only model also fares well, but its Sharpe and maximum draw-down are still behind the sentiment-enhanced version.

Cross-Window Takeaway

Across all horizons, adding Reddit sentiment improves absolute and risk-adjusted performance, while never increasing draw-down. Trade frequency either falls (10-day) or rises only marginally (15-day, 20-day). This implies that crowd mood harvested from Reddit is an economically valuable complement to price/volume features – at least for the 2024 Bitcoin regime examined here.

5.2.2 MACD Gate-Classifiers

10-day window

The Random Forest gate rides above every other curve for most of the sample and closes the period with a far larger gain than the plain MACD. Its 2018 and 2022 draw-downs are also noticeably shallower. Back in Section 5.1.2, we saw that the 10-day forest posted the highest F_1 of all gates and kept both FPR and FNR close to 40%. That balance produced fewer but higher quality trades, which shows up here as a smoother and steeper equity line. In short, a 10-day sentiment window captures the rapid mood swings that often precede price changes.

15-day window

At the 15-day horizon, the forest doubles the baseline's return – opposite to what we observed in Track 1. Moderate smoothing appears to strip-out some high-frequency noise, letting the

forest's non-linear splits focus on the stronger sentiment signals. The result is a clean boost relative to both the baseline and the simpler gates.

20-day window

When the sentiment window is extended to 20 days, the forest and the baseline finish more or less even, while Ridge-Logit and Naïve Bayes remain stuck at the bottom. Section 5.1.2 already showed recall slipping for the forest at this setting, which is also confirmed by the equity curves due to the filtering of too many good trades. Heavy smoothing dulls the cues that distinguish strong and weak MACD signals at the shorter look-back windows.

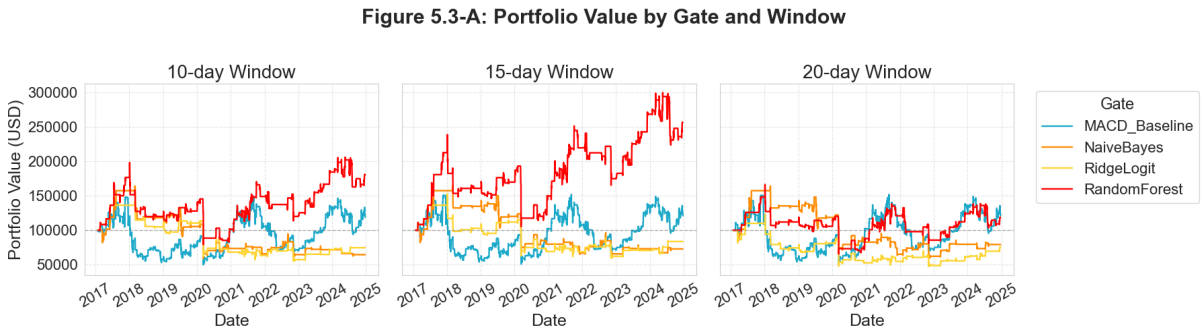

Figure 5.6: Simulated portfolios for the gates with the different window sizes

Takeaways

We observe that both the Naïve Bayes and the Ridge-Logit post negative results across all windows despite respectable hit rates 5.2.2. Their higher false-negative rates (Section 5.1.2), mean they sit out most of the big bull-legs and only capture small winners. Their conservatism leads to too many missed-out opportunities. Furthermore, we observe that sentiment needs a flexible reader as crowd mood is no silver bullet on its own, though it can increase economic value of the simplified MACD rule, when utilized by the right model. Longer windows tend to erase the cues stored in shorter ones, leading to a fading edge. Therefore, the information value of sentiment is both time-scale dependent and model dependent.

Table 5.2: Performance Metrics – Track 2 (MACD Gates)

Window	Model	Trades	HitRate	MeanNet	Sharpe	MaxDD	TotalReturn
10d	Ridge-Logit	43	0.4884	0.0008	-0.2045	-0.6130	-0.2541
10d	Naïve Bayes	45	0.5111	-0.0008	-0.2270	-0.6398	-0.3551
10d	Random Forest	76	0.5789	0.0150	0.2429	-0.6005	0.8081
15d	Ridge-Logit	45	0.5333	0.0036	-0.1280	-0.6105	-0.1636
15d	Naïve Bayes	45	0.5111	0.0019	-0.1670	-0.6158	-0.2723
15d	Random Forest	77	0.5844	0.0194	0.3840	-0.5596	1.5639
20d	Ridge-Logit	50	0.5200	0.0018	-0.1595	-0.6146	-0.2433
20d	Naïve Bayes	51	0.5098	0.0039	-0.1140	-0.6364	-0.2075
20d	Random Forest	74	0.5270	0.0095	0.0640	-0.6064	0.1857
–	MACD Baseline	152	0.5395	0.0081	0.1656	-0.6627	0.2095

5.3 Statistical Inference

To understand sampling uncertainty around the Sharpe ratio gain from sentiment, I generated 10 000 IID-bootstrap[1] resamples of the Sharpe-ratio difference

$$\Delta SR = SR_{sentiment} - SR_{indicator}$$

for each look-back window. Figure 5.8 displays the resulting sampling distributions, while Table 5.3 reports the 95% percentile confidence intervals, the bootstrap probabilities ($P(\Delta SR > 0)$), and one-sided Wilcoxon signed-rank p-values.

The bootstrap means are positive in every panel, and the posterior-like probabilities range from 0.75 to 0.88, indicating that the sentiment-enhanced model beats the pure indicator model roughly four out of five resamples.

All three 95% confidence intervals include zero, thus the gain is not significant at the 5% level. The Sharpe ratios are notoriously noisy, thus detecting an edge larger than a few tenths requires either many more trades or a significantly larger effect size.

The non-parametric Wilcoxon test - performed to introduce robustness - yields one-sided p-values of 0.17, 0.20, and 0.055 for the 10-day, 15-day, and 20-day windows, respectively. The 20-day result borders significance, providing moderate evidence of an edge.

Lastly, applying the Benjamini-Hochberg false-discovery-rate procedure with $q = 0.10$, leaves all three hypotheses un-rejected. The relatively liberal choice of $q = 0.10$ balances power against Type-I errors; a stricter threshold would only reinforce the non-significant outcome.

Overall, the point estimates and high $P(\Delta SR > 0)$ values consistently favor the sentiment filter, yet the formal tests cannot rule out a zero Sharpe gain. Hence, the evidence added for economic value is moderate but not decisive.

[1] For the fulfillment of the IID-assumptions see the Appendix (A.2)

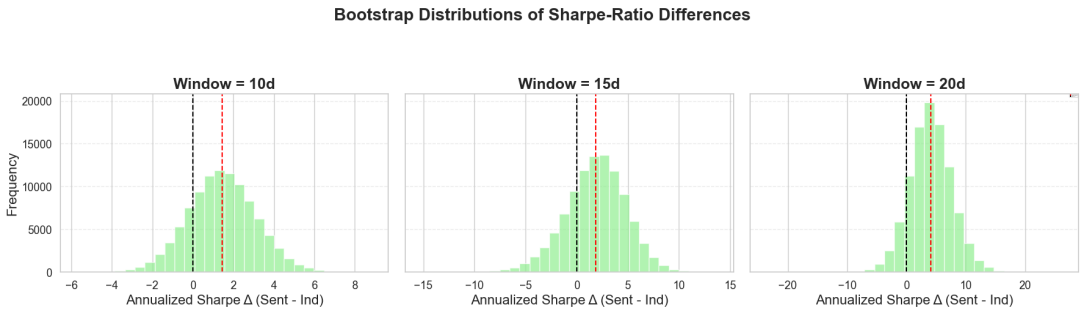

Figure 5.7: Bootstrap of the Sharpe ratio difference of the models across the different windows

Table 5.3: Bootstrap Confidence Intervals and Wilcoxon Test Results

Window	**CI Lower**	**CI Upper**	$P(\Delta > 0)$	**Wilcoxon** p
10 d	−1.840	4.652	0.801	0.170
15 d	−4.198	7.193	0.753	0.202
20 d	−2.754	11.370	0.879	0.055

Table 5.4: Benjamini–Hochberg–Adjusted One-Sided p-Values ($q = 0.10$)

Window	$\mathbf{p}_{\text{one-sided}}$	$\mathbf{p}_{\text{BH}}$	**Reject** H_0**?**
10 d	0.199	0.247	No
15 d	0.247	0.247	No
20 d	0.121	0.247	No

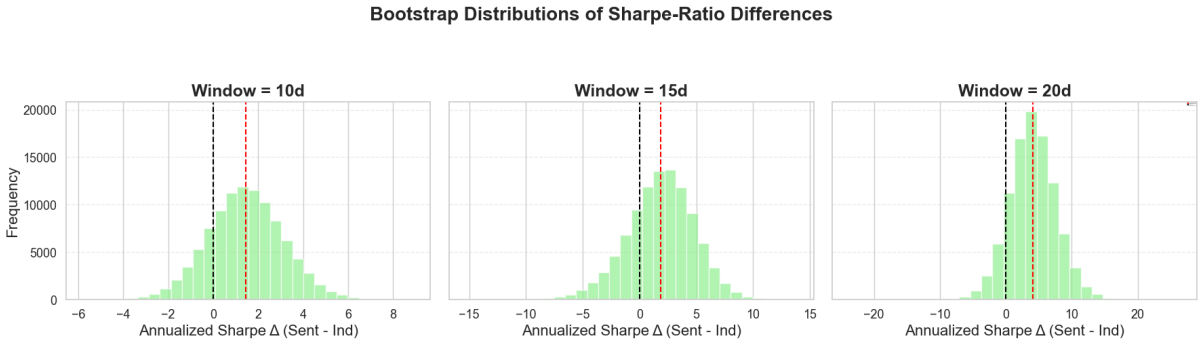

Figure 5.8: Bootstrap Distributions of the Sharpe Ratio difference (B = 10000)

5.4 The Effect of Sentiment

Track 1 relies entirely on the Random-Forest classifier to decide whether a five-day trade is worthwhile; there is no mechanical trigger such as a MACD crossover. To understand why the 20-day sentiment-enhanced forest delivers the strongest economic results, I inspect SHAP values for the 2024 test set (Figures 5.9 5.10 5.11) and draw my main conclusion based on it, as well as I give insight into the effect across the different look-back lengths.

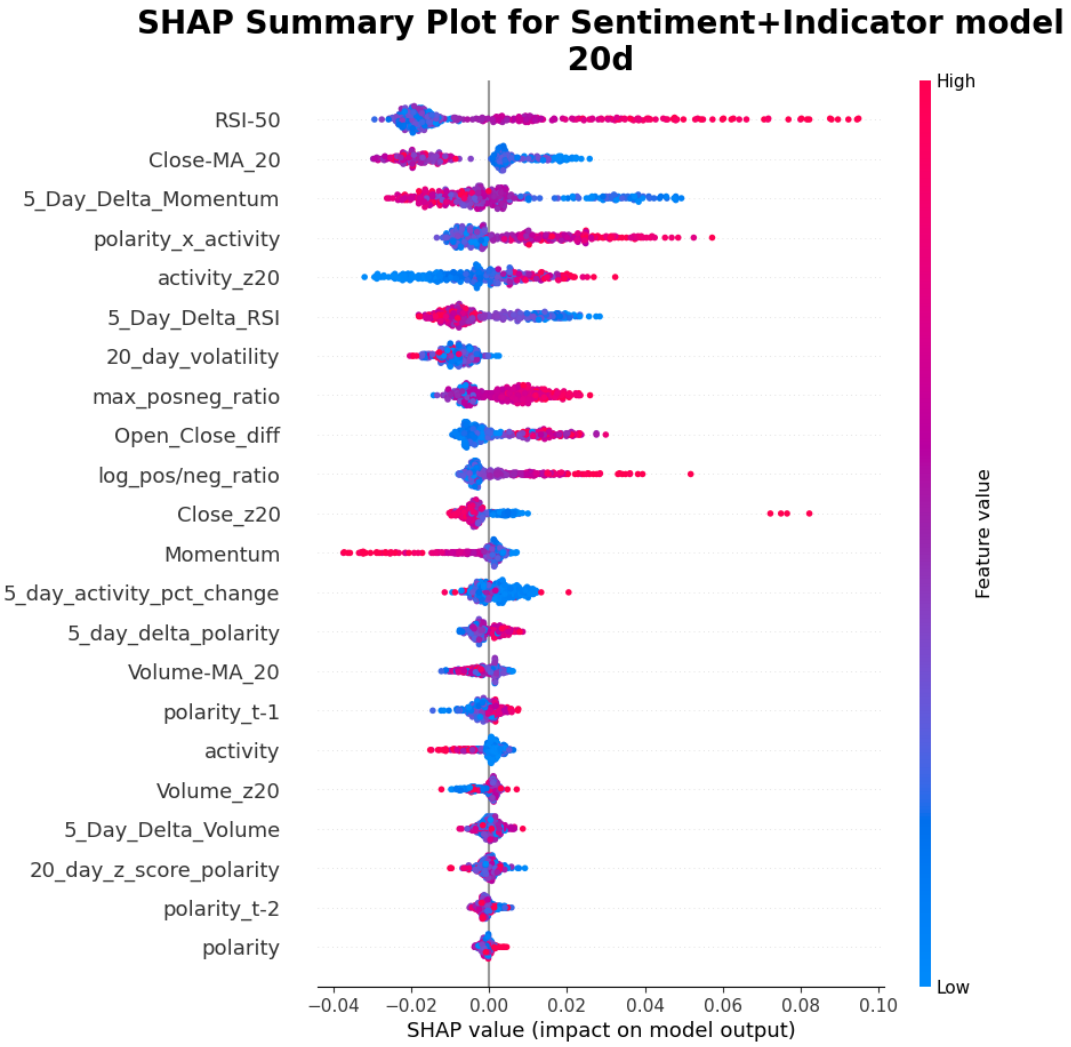

Figure 5.9: Beeswarm plot of the SHAP values for the 20-day window

The bee-swarm plot (Figure 5.9) ranks features by their average absolute impact on the model's log-odds. `RSI-50` remains the single most influential variable, reflecting its long-standing reputation as a stand-alone momentum gauge in financial markets [34]. A sentiment feature breaks into the top features consistently (Figures A.3 A.4): the interaction term `polarity_x_activity`. Pure tone measures such as raw polarity appear further down the list, suggesting that at a 20-day horizon engagement intensity weighted sentiment outranks pure sentiment and engagement level.

Figure 5.10 focuses on `polarity_x_activity`. SHAP contributions are essentially zero — or even negative — until the interaction exceeds a threshold of about 1.0; beyond that point they rise almost linearly, reaching roughly six basis points at the upper end of the observed range. Coloring each point by `RSI-50` reveals no clear gradient, indicating that the interaction is largely orthogonal to RSI. In practical terms, the forest opens a position only when the price is believed to be strong and people are actively discussing it. This dual filter explains the model's behavior documented in Section 5.2.1: it fires only 14 times yet achieves an 86% hit rate and a

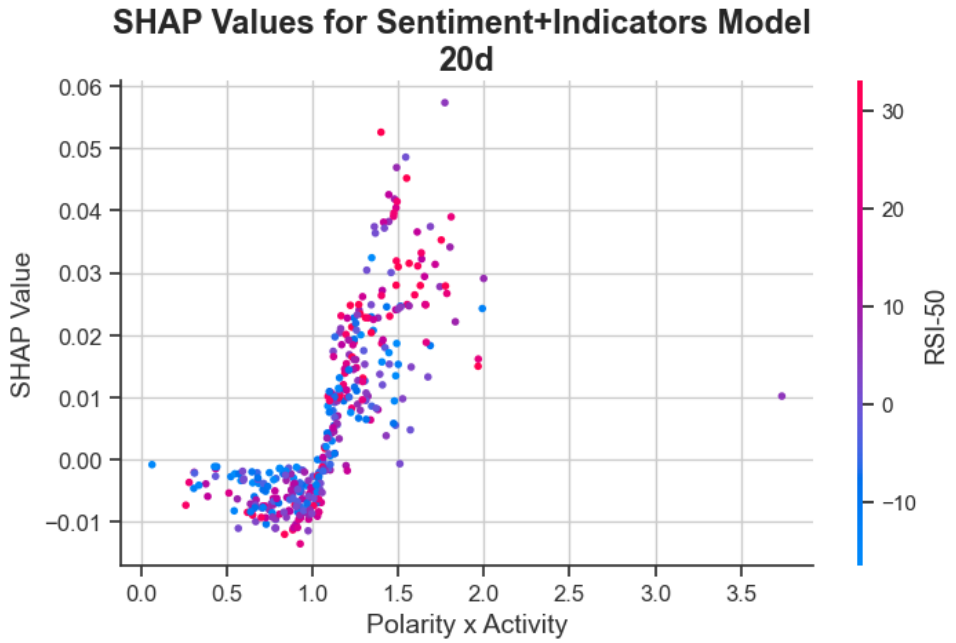

Figure 5.10: SHAP values for `polarity_x_activity`, colored by `RSI-50`

Sharpe-ratio above 2.3, because many technically plausible set-ups are vetoed for lack of social confirmation.

Figure 5.11 reverses the view, plotting SHAP against `RSI-50` and coloring by `polarity_x_activity`. Positive RSI values lift SHAP sharply; sub-zero values depress it, reflecting pure trend-following rather than the classical over-bought/over-sold interpretation [34]. When both RSI, and the interaction term are high, points cluster in the extreme upper-right, confirming that the two features complement rather than substitute each other. This behavior aligns with evidence that Bitcoin trends persist longer than those in large-cap equities [35] and with De Long et al.'s noise-trader framework, which links bursts of one-sided sentiment to short-run momentum [1].

Yesterday's sentiment (`polarity t{1`) still contributes positively, although its mean SHAP value is modest and its rank slips to twelfth. The observed influence is consistent with Chen et al., who observe that yesterday's sentiment has a significant effect on next day's returns [22]. It also reflects the earlier observation that sentiment helps most at the 10-day window and then gradually falls behind longer-term technical cues.

Taken together, the SHAP-analysis shows that crowd mood acts as a selective amplifier rather than an autonomous trigger. It boosts confidence in technically attractive, high-buzz regimes and suppresses trades when price looks healthy but the conversation is muted. By screening signals through this behavioral lens, the 20-day sentiment forest preserves upside while keeping draw-downs shallow, thereby translating its statistical edge into the superior risk-adjusted returns documented in the back-test.

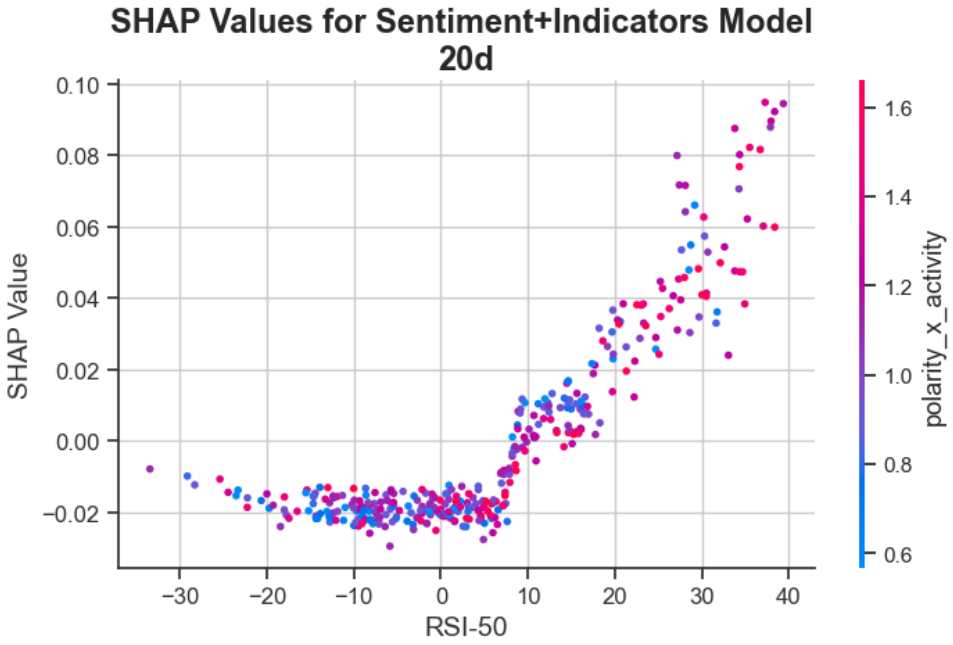

Figure 5.11: SHAP values for `RSI-50`, colored by `polarity_x_activity`

Chapter 6

Conclusion

6.1 Key Findings

Sentiment consistently improves Random-Forest timing – especially at the 20-day range. Across every look-back window, sentiment features raise precision and trim the surge in false negatives that handicaps the 15- and 20-day indicator-only models. At 20-days, the sentiment forest executes only 14 trades, yet delivers an 83% total return, an 86% hit rate, Sharpe = 2.30 and MaxDD = -6.6%, outperforming both the indicator-only model and the buy-and-hold. SHAP analysis shows that `polarity_x_activity` acts as a "buzz filter", green-lighting only those momentum setups that coincide with intense, upbeat discussion, thus boosting high hit rate without flooding the portfolio with trades.

MACD signals benefit from a flexible sentiment gate, but not from a linear or naive one. The depth-3 Random Forest gate roughly halves the trade count while nudging precision and F_1 upward across all windows. The Ridge-Logit and Naive Bayes block profitable as well as unprofitable crossovers, leading to negative equity curves in every horizon. This result underscores that sentiment is no silver bullet, it adds value only when the model can exploit non-linear interactions.

Statistical evidence is mixed: point estimates favor sentiment, but sampling noise is high. The IID-bootstrap probabilities that the sentiment Sharpe-ratio exceeds the baseline range from 0.75 to 0.88, but the 95% confidence intervals include zero and Wilcoxon p-values stay well above 0.05 for the 10- and 15-day windows. The Benjamini-Hochberg correction leaves all three Sharpe-gain hypotheses un-rejected, thus the economic edge is promising, but not decisive.

Engagement-weighted sentiment – not raw polarity – drives the incremental alpha. In all forests `polarity_x_activity` ranks among the top features (behind `RSI-50`). SHAP

scatter plots show a hinge at engagement-intensity (about 1.0): only when crowd tone and activity are simultaneously high does the model's log-odds increase materially, explaining why the 20-day strategy executes few trades, yet posts an 86% hit rate. This pattern echoes noise-trader theory – periods of one-sided, high-attention sentiment exert short-run price pressure [1].

Time-scale and model flexibility are critical.
For Random Forests, short windows capture fast decaying sentiment cues, long windows align with regime-level shifts, and the mid-range horizons blur both. In the MACD-gating setup, the 15-day window performs best because moderate smoothing removes high-frequency noise while leaving enough variation for the gate to separate strong from weak crossovers. Non-linear learners are required to exploit most of the sentiment edge. These results caution against one-size-fits all approaches and present sentiment rather as a complementary, time-scale dependent signal whose value emerges only when paired with a flexible model.

6.2 Implications

The evidence fits comfortably within the behavioral-finance literature and showcases the practical limits of a strict Efficient Market Hypothesis (EMH). In line with Barberis & Thaler, the results suggest that some market participants are not fully rational, and that limits to arbitrage prevent prices from instantaneously correcting their sentiment-driven deviations [9]. Our sentiment-enhanced model improves prediction accuracy and lifts risk-adjusted returns, implying that investor mood does leak into prices in an exploitable way.

Yet, the edge is small and statistically fragile. That nuance echoes De Long et al.'s noise trader risk framework: sentiment creates local mis-pricing, but exploiting them is costly and risky, so the extra Sharpe we observe weakens once we introduce sampling variability [1]. This means that the results challenge the strictest form of EMH, while agreeing with its weaker version: there is some predictability in the market, but due to the strong competition, it starts to fade quickly.

Within the cryptocurrency literature, our results reinforce recent findings: Mai et al., Ider & Lessmann and Gurgul et al. show that social media and news sentiment have a strong predictive power for Bitcoin and other coins [14] [19] [15]. I replicate this pattern: sentiment consistently boosts classification metrics and provides moderate evidence of higher Sharpe-ratios.

The study also contributes to the debate on technical analysis. Early optimism [28] held that simple moving average rules sufficed for an edge, later Bajgrowicz & Scaillet demonstrated that the apparent profits evaporate once one controls for externalities [29]. Our indicator-only baseline under-performs at longer horizons, but regains relevance once filtered through sentiment. Hence technical analysis is still viable, it just needs extra context in the form of behavioral signals to remain competitive.

6.3 Limitations and Further Research

This study intentionally keeps the trading framework simple, so the reader can see what sentiment adds – yet that simplicity also sets the boundaries of what I can claim. A main shortcoming lies in how I design the signals. I only let the classifiers give long signals for a fixed-holding day period. That choice hides at least two additional sources of information: (i) bearish sentiment that could trigger short- or flat positions, and (ii) intra-holding updates – warnings that arrive on day 2 or day 3 are ignored because the trade is already locked in. A multi-directional model paired with dynamic exits would tell us whether sentiment can do more than pick entry points. I also used a stripped-down, long-only MACD strategy for Track 2; a bi-directional MACD might close part of the gap and give us a clearer picture about the incremental value of sentiment.

On the strategy mechanics side, the back-test has no stop-loss, no position sizing and – importantly – no plan for idle capital. When the classifier predicts flat (or downturn) periods it leaves cash standing. Exploiting such periods could turn passive times into profitable ones as well, further rewarding the stronger models, and thus highlighting their superior performance. Likewise, even modest stop-losses would have intervened in some of the draw-down periods of the different strategies, improving their metrics.

The third limitation is data quality: the current sentiment features are CryptoBERT-scores derived from Reddit. Cryptocurrency-related conversations happen on other media platforms like Twitter, Discord and - importantly - news channels. These could enrich our sentiment dataset, and perhaps lead to more conclusive results. Sentiment could also be mixed with option-implied volatility, GARCH forecasts, or stochastic volatility models to capture regimes where emotion and uncertainty move together.

Finally, the evaluation framework itself can be improved and further validated. A walk-forward live-test would reveal whether the edge persists once the model is out in the real world. One could also dive deeper into the mechanics of the sentiment analysis process and try to analyze which phrases and interactions are driving model decisions.

In summary, sentiment clearly offers extra information, but the edge is small and fragile. Broader data, smarter position management, and tougher out-of-sample tests should be further experimented with to turn this statistically suggestive result into a strategy that exploits real financial markets.

6.4 Closing Remarks

This thesis set out to test whether crowd mood harvested from Reddit adds economic value to classical technical indicators in predicting short-term Bitcoin returns. By integrating sentiment into both machine learning predictors and gating filters, I show that social conversations can

sharpen timing, while reducing risk. Although the study is biased upwards by a bullish 2024 market, the transparent evaluation and research design provide a robust template for extending sentiment trading to broader asset classes. In an era where information flows are both noisy and high-volume, fusing traditional signals with crowd emotions appears not just feasible, but financially rewarding.

Bibliography

[1] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise Trader Risk in Financial Markets," *The journal of political economy.*, vol. 98, Aug. 1990. Place: Chicago, Ill. : Publisher: University Press of Chicago,.

[2] J. M. Keynes, *The General Theory of Employment, Interest and Money*. London: Macmillan, 1936.

[3] M. BAKER and J. WURGLER, "Investor Sentiment and the Cross-Section of Stock Returns," *The journal of finance.*, vol. 61, Aug. 2006. Place: [Malden, Mass.] : Publisher: Published by Blackwell Publishers for the American Finance Association.

[4] M. Baker and J. Wurgler, "Investor Sentiment in the Stock Market," *The journal of economic perspectives : a journal of the American Economic Association.*, vol. 21, Apr. 2007. Place: Nashville, TN : Publisher: American Economic Association,.

[5] S. C. Long, B. Lucey, Y. Xie, and L. Yarovaya, ""I just like the stock": The role of Reddit sentiment in the GameStop share rally," *Financial Review*, vol. 58, pp. 19–37, Feb. 2023.

[6] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of finance (New York)*, vol. 25, no. 2, pp. 383–, 1970.

[7] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica : journal of the Econometric Society.*, vol. 47, Mar. 1979. Place: Chicago, Ill. : Publisher: Econometric Society, the University of Chicago.

[8] W. F. M. De BONDT and R. THALER, "Does the Stock Market Overreact?," *The journal of finance.*, vol. 40, July 1985. Place: [Malden, Mass.] : Publisher: Published by Blackwell Publishers for the American Finance Association.

[9] N. Barberis and R. Thaler, "Chapter 18 A survey of behavioral finance," in *Financial Markets and Asset Pricing*, vol. 1 of *Handbook of the Economics of Finance*, pp. 1053–1128, Elsevier, 2003. ISSN: 1574-0102.

[10] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[11] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[12] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," Aug. 2019. arXiv:1908.10063 [cs].

[13] M. Kulakowski and F. Frasincar, "Sentiment classification of cryptocurrency-related social media posts," *IEEE Intelligent Systems*, vol. 38, no. 4, pp. 5–9, 2023.

[14] F. Mai, J. Shan, Q. Bai, S. Wang, and R. Chiang, "How does social media impact bitcoin value? a test of the silent majority hypothesis," *Journal of Management Information Systems*, vol. 35, pp. 19–52, 01 2018.

[15] V. Gurgul, S. Lessmann, and W. K. Härdle, "Deep learning and nlp in cryptocurrency forecasting: Integrating financial, blockchain, and social media data," *International Journal of Forecasting*, 2025.

[16] Academic Torrents, "Academic torrents: Distributed data sharing for researchers." https://www.academictorrents.com. Accessed: 2025-03-11.

[17] Yahoo Finance, "Yahoo finance - cryptocurrency market data." https://finance.yahoo.com. Accessed: 2025-03-11.

[18] R. Aroussi, "yfinance: Download market data from yahoo! finance's api." https://github.com/ranaroussi/yfinance, 2019. Accessed: 2025-03-11.

[19] D. Ider and S. Lessmann, "Forecasting cryptocurrency returns from sentiment signals: An analysis of bert classifiers and weak supervision," *arXiv preprint arXiv:2204.05781*, 2022.

[20] H. Arian, D. N. Mobarekeh, and L. Seco, "Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment," *Knowledge-Based Systems*, vol. 305, p. 112477, 2024.

[21] D. H. Bailey, J. M. Borwein, M. L. De Prado, and Q. J. Zhu, "Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance," *Notices of the AMS*, vol. 61, no. 5, pp. 458–471, 2014.

[22] C. Chen, R. Despres, L. Guo, and T. Renault, "What makes cryptocurrencies special? investor sentiment and return predictability during the bubble," *SSRN Electronic Journal*, 01 2019.

[23] S. Nasekin and C. Y.-H. Chen, "Deep learning-based cryptocurrency sentiment construction," *Digital Finance*, vol. 2, no. 1, pp. 39–67, 2020.

[24] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500," *European Journal of Operational Research*, vol. 259, no. 2, pp. 689–702, 2017.

[25] J. Kalyani, P. Bharathi, P. Jyothi, *et al.*, "Stock trend prediction using news sentiment analysis," *arXiv preprint arXiv:1607.01958*, 2016.

[26] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[27] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[28] W. Brock, J. Lakonishok, and B. LeBaron, "Simple technical trading rules and the stochastic properties of stock returns," *The Journal of finance*, vol. 47, no. 5, pp. 1731–1764, 1992.

[29] P. Bajgrowicz and O. Scaillet, "Technical trading revisited: False discoveries, persistence tests, and transaction costs," *Journal of Financial Economics*, vol. 106, no. 3, pp. 473–491, 2012.

[30] B.-K. Kang, "Improving macd technical analysis by optimizing parameters and modifying trading rules: Evidence from the japanese nikkei 225 futures market," *Journal of Risk and Financial Management*, vol. 14, no. 1, 2021.

[31] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[33] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[34] J. W. J. Wilder, *New Concepts in Technical Trading Systems*. Greensboro, NC: Trend Research, 1978.

[35] O. Borgards, "Dynamic time series momentum of cryptocurrencies," *The North American Journal of Economics and Finance*, vol. 57, p. 101428, 2021.

Appendix A

Appendix

A.1 Code Availability

The code and data used for this analysis can be found on GitHub.

A.2 Assumptions for IID-Bootstrap

The IID-Bootstrap assumes that the resamples come from independent and identically distributed original sample. For this I perform the Ljung-Box Test (independence) and split the returns in half in time and check a 2-Sample Kolmogorov-Smirnov Test (identically distributed) A.2. I do not reject any hypothesis, thus I do not have any reason to believe that the IID-Bootstrap will bias the results.

Table A.1: KS Split–Half and Ljung–Box Tests on 5-Day Trade Returns

Window	**Model**	n	**KS stat**	**KS** p	**LB stat**	**LB** p
10d	Indicator	57	0.1564	0.8045	1.4425	0.9196
10d	Sentiment	48	0.2083	0.6860	2.0782	0.8382
15d	Indicator	17	0.5139	0.1732	5.2852	0.3821
15d	Sentiment	34	0.2353	0.7506	7.5422	0.1833
20d	Indicator	13	0.3810	0.6224	3.3779	0.6419
20d	Sentiment	14	0.2857	0.9627	9.1205	0.1044

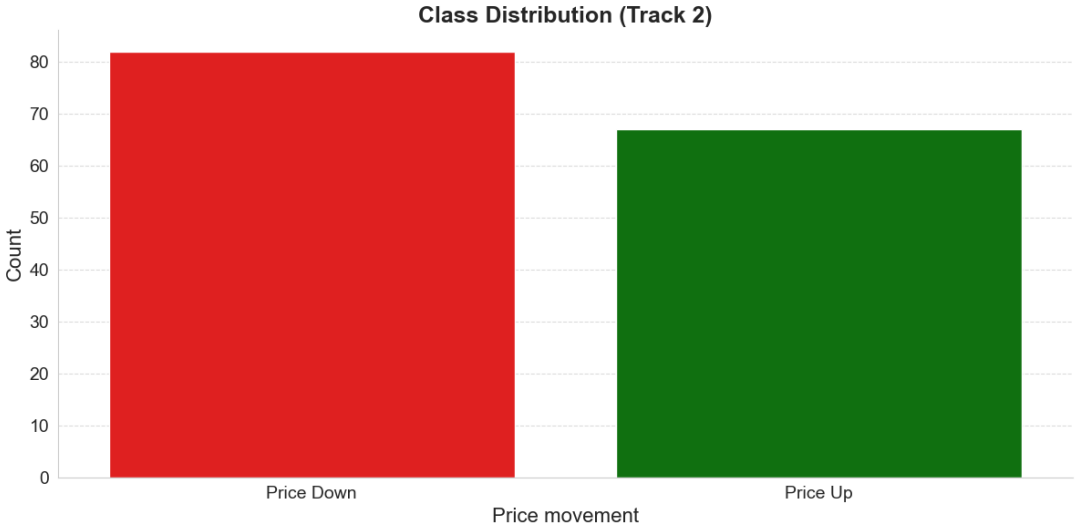

Figure A.2: Distribution of the classes in the second model setup

A.3 Extra plots

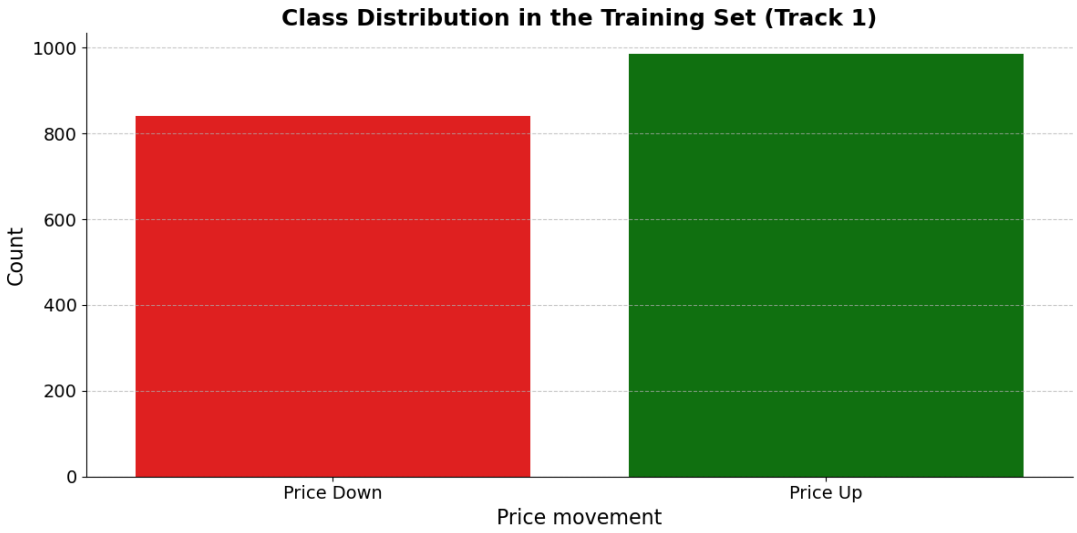

Figure A.1: Distribution of the classes in the first model setup

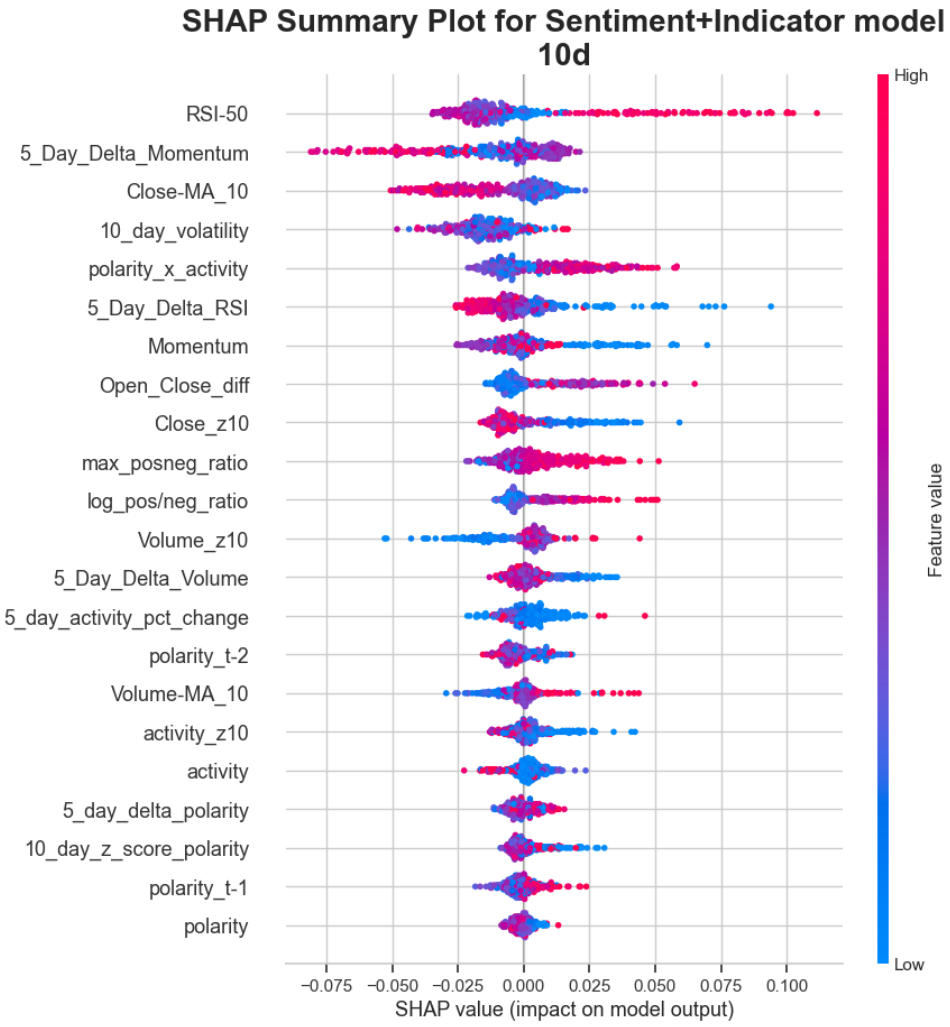

Figure A.3: Beeswarm plot of the SHAP values for the 10-day window

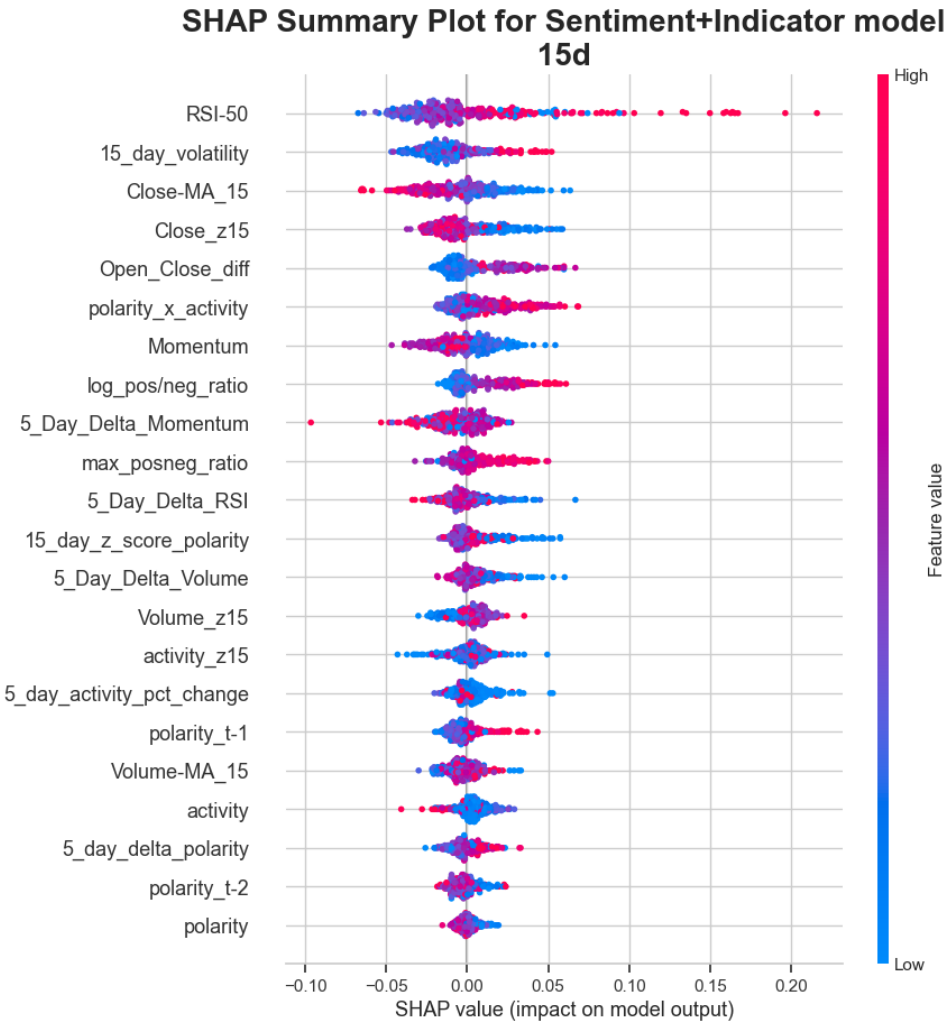

Figure A.4: Beeswarm plot of the SHAP values for the 15-day window