Mining Market Sentiment: A Comparative Analysis of Domain-Specific Language Models on Cryptocurrency News for Market Forecasting

Karl Saliba, Michel Saliba*

May 27, 2026

Abstract

The volatile nature of cryptocurrency markets presents a unique challenge for predictive forecasting, where media sentiment often drives rapid price fluctuations. This study presents a comprehensive sentiment analysis pipeline evaluating five distinct natural language processing models—TextBlob, CryptoBERT, FinBERT, and Qwen2.5-7B-Instruct; on a dataset of historical cryptocurrency news spanning 2021 to 2023. We first conduct a statistical inter-comparison of the models to evaluate their baseline agreement and expose inherent, domain-specific biases in how they assess crypto-economic terminology. Subsequently, we benchmark the aggregated sentiment signals generated by each model against hard financial market data to evaluate their real-world predictive efficacy. Our findings reveal that while apex assets like Bitcoin resist isolated sentiment forecasting, open-weight generative models such as Qwen2.5-7B-Instruct demonstrate a distinct predictive edge for Ethereum and Altcoin directions out-of-the-box. Yielding an average accuracy of 53%; with a 95% confidence interval predominantly distributed above the 50% random-chance baseline; these results prove that parameter-efficient architectures offer a highly viable, computationally accessible solution for crypto-economic analysis without the need for bespoke retraining.

1 Introduction

Financial markets are inherently volatile, but cryptocurrency markets exhibit an extreme degree of rapid and unpredictable price fluctuation [1]. Due to the lack of traditional valuation metrics, digital asset prices are heavily driven by public perception, investor behavior, and media narratives. Consequently, sentiment analysis of financial news has emerged as a powerful tool for market forecasting, with recent literature demonstrating its efficacy in extracting predictive signals from textual data [2].

Despite the proven utility of news-based market prediction, acquiring comprehensive historical datasets of cryptocurrency news remains a challenge. Researchers frequently encounter data scarcity due to URL expiration, link rot, and the ephemeral nature of digital media publications [3]. This lack of persistent, large-scale textual data makes it computationally expensive and difficult to train specialized Natural Language Processing (NLP) models from scratch. As a result, the financial technology sector often relies on pre-existing, "open-box" models. However, while some of these open-weight models were fine-tuned on traditional financial corpora, their direct applicability and accuracy when exposed to the unique, hyper-specific lexicon of the cryptocurrency domain remain uncertain.

The objective of this study is to evaluate whether existing open-weight models—specifically variants of BERT and Llama architectures—can be effectively utilized for cryptocurrency market prediction without requiring ground-up training. We conduct an inter-comparison of five distinct NLP models to assess their baseline agreement when scoring cryptocurrency news. Subsequently, we benchmark the sentiment signals generated by each model against actual cryptocurrency market data spanning from 2021 to 2023. This specific time-frame was selected as it represents a period of extreme macroeconomic shifts, geopolitical stress, and unprecedented boom-bust volatility within the crypto-economy [4, 5], providing a rigorous testing environment for predictive efficacy.

The remainder of this paper is structured as follows. Section 2 details the methodology, outlining the selected NLP models, the dataset acquisition process, and the data analysis techniques employed.

*michel.saliba@epfl.ch

Section 3 is divided into two parts: the first presents the results of the inter-model agreement analysis, while the second provides a direct predictive comparison of the sentiment models against historical market data. Finally, Section 4 offers concluding remarks and directions for future research.

2 Methodology

This section outlines the technical framework of our study, detailing the selection of sentiment analysis models, the provenance of our historical datasets, and the statistical metrics used to evaluate model performance against market movements.

2.1 Sentiment Analysis Models

To ensure a robust comparison, we selected five models representing different architectural approaches to Natural Language Processing (NLP), ranging from specialized encoder-only models to large-scale generative decoders.

2.1.1 Transformer-Based Encoders: FinBERT and CryptoBERT

We utilized two domain-specific variants of the BERT architecture [6]; the architecture of the model is shown in Figure 1.

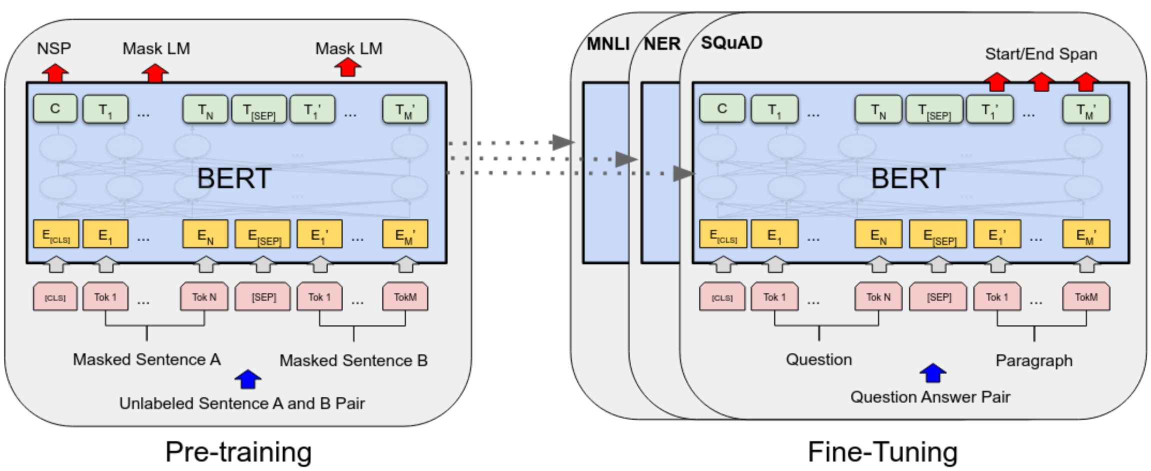


Figure 1: The BERT methodology maintains architectural consistency across both its pre-training and fine-tuning phases, differing only in the task-specific output layers. When adapting to various downstream tasks, the model initializes with the same pre-trained weights, and all parameters undergo updates during the fine-tuning process. Additionally, input formatting relies on specialized tokens: `[CLS]` is prepended to every input sequence, while `[SEP]` functions as a boundary marker between distinct text segments.[6]

- **FinBERT:** A model pre-trained on a large-scale financial corpus to better understand nuances in financial communication [7, 8].
- **CryptoBERT:** A Roberta model specifically fine-tuned on social media and news data from the cryptocurrency sector to capture industry-specific slang and sentiment [9, 10].

2.1.2 Dense Generative Architectures: Qwen2.5

To evaluate an alternative foundational architecture, we implemented Qwen2.5-7B-Instruct, a capable open-weight model recognized for its advanced instruction-following capabilities and strong technical reasoning performance [11]. The specific architecture utilized in this study operates as a *dense* model. In a dense neural network, every parameter within the model is activated and utilized during the forward pass for every single token processed. This is in contrast to sparse architectures, such as

Mixture of Experts (MoE), which dynamically route tokens to specialized sub-networks. The dense nature of the Qwen2.5-7B-Instruct model ensures a comprehensive application of its pre-trained 7-billion parameter knowledge base to every cryptocurrency news headline. This structural characteristic provides a stable, exhaustive, and highly predictable computational pathway for our sentiment analysis pipeline, guaranteeing that the model's full reasoning capacity is applied to each linguistic nuance.

2.1.3 Lexicon-Based Baselines: TextBlob

To establish a foundational, deterministic baseline for our sentiment analysis, we incorporated TextBlob, a widely utilized lexicon-based NLP library built upon the Natural Language Toolkit (NLTK) and Pattern [12]. Unlike transformer-based models that dynamically compute dense contextual embeddings, TextBlob operates via a predefined linguistic rule set. It analyzes textual inputs to return a tuple containing two primary metrics: polarity and subjectivity. Polarity is quantified as a continuous float score ranging from -1.0 (highly negative) to 1.0 (highly positive), while subjectivity ranges from 0.0 (strictly objective) to 1.0 (highly subjective). By calculating the aggregate polarity of recognized lexical tokens, TextBlob provides a transparent, non-contextual sentiment baseline. Including this traditional approach is crucial, as it allows us to precisely quantify the performance delta—and potential contextual advantages—achieved by deploying computationally heavier, machine-learning-driven architectures.

2.2 Data Acquisition

The data used in this study is bifurcated into textual sentiment data and quantitative asset pricing data.

2.2.1 Sentiment Dataset

Historical cryptocurrency news spanning 2021 to 2023 was sourced from the "Crypto News" dataset curated by oliviervha via Kaggle [13]. This dataset provides the foundational textual inputs for our NLP pipeline, specifically utilizing timestamps, headlines, and article snippets to chronologically map market sentiment. To align this textual data with our target market assets, we filtered the corpus by its pre-assigned subject tags: articles categorized under "ethereum" were utilized to track ETH prices, "blockchain" were utilized for BTC, and "altcoin" was isolated to track Dogecoin (DOGE).

2.2.2 Asset Pricing Data

Historical market data for Bitcoin (BTC), Ethereum (ETH), and DOGE coin (DOGE) were retrieved via the Yahoo Finance API using the `yfinance` Python library [14]. To align with our aggregated daily sentiment signals, we extracted the daily adjusted closing prices and trading volumes for each respective asset.

2.3 Data Analysis and Evaluation Metrics

Our analysis pipeline processes raw news into actionable signals using the following:

2.3.1 Market Direction Calculation

Market movement is defined as a binary classification where the direction D_t for a given day t is determined by:

$$D_t = \begin{cases} 1 & \text{if } P_{t+1} > P_t \\ 0 & \text{if } P_{t+1} \leq P_t \end{cases} \tag{1}$$

where P_t represents the closing price.

2.3.2 Inter-Model Agreement

To measure the reliability and domain-specific bias between our chosen models, we utilize confusion matrices alongside Cohen's Kappa coefficient (κ) [15]. While a confusion matrix provides a granular, cross-tabular visualization of exactly where models align or diverge across specific sentiment classes (positive, neutral, negative), Cohen's Kappa provides a robust, single-metric statistical summary of this inter-rater reliability. The Kappa coefficient is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

where p_o is the relative observed agreement among the models, and p_e is the hypothetical probability of chance agreement. This combined approach allows us to not only visualize specific directional biases (e.g., if one model skews overly bullish), but also to mathematically quantify if two models are truly extracting the same underlying sentiment, or if their consensus is merely incidental.

2.3.3 Sentiment Aggregation and Signal Generation

To benchmark the NLP models against binary market movements, it was necessary to map their categorical outputs (Positive, Neutral, Negative) into a quantifiable daily sentiment signal. First, each individual news article i published on day t was assigned a discrete numerical polarity score $s_{i,t}$ based on the model's classification:

$$s_{i,t} = \begin{cases} 1 & \text{if Positive} \\ 0 & \text{if Neutral} \\ -1 & \text{if Negative} \end{cases} \tag{3}$$

For any given day t containing N_t news articles, the aggregated daily sentiment score S_t was calculated as the arithmetic mean of all individual article polarities [16]:

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} s_{i,t} \tag{4}$$

Finally, this continuous daily sentiment score was converted into a predicted market direction $\hat{D}_t$. To align with our prior definition of actual market movement (where 1 indicates a price increase and 0 indicates a decrease or stagnation), the predictive signal was formulated as:

$$\hat{D}_t = \begin{cases} 1 & \text{if } S_t > 0 \quad \text{(Bullish Signal)} \\ 0 & \text{if } S_t \leq 0 \quad \text{(Bearish/Neutral Signal)} \end{cases} \tag{5}$$

By comparing the predicted direction $\hat{D}_t$ against the actual market direction D_t, we were able to calculate the absolute predictive accuracy of each model for the selected assets.

2.3.4 Predictive Accuracy and Confidence Intervals

To evaluate the predictive efficacy of the aggregated sentiment signals against actual market movements, we utilized standard binary classification metrics. The primary evaluation metric is Classification Accuracy, which calculates the proportion of correctly predicted market directions (Up vs. Down):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP and TN represent the true positives and true negatives (correct predictions of upward and downward movements, respectively), while FP and FN represent the false positives and false negatives.

To ensure statistical robustness and account for sample variance across the testing period, we also calculated the Standard Error (SE) for each model's accuracy:

$$SE = \sqrt{\frac{p(1-p)}{n}} \tag{7}$$

where p is the observed accuracy proportion and n is the total number of trading days analyzed. From this, we constructed 95% confidence intervals by applying a multiplier of $1.96 \times SE$. This allowed us to establish clear error margins and rigorously determine if the performance differences between the open-box architectures were statistically significant.

3 Results and Discussion

This section presents the empirical findings of our sentiment analysis pipeline. We first evaluate the inter-model agreement and inherent biases across the five distinct NLP architectures. Subsequently, we benchmark their aggregated sentiment signals against actual cryptocurrency market movements to assess predictive efficacy.

3.1 Inter-Model Agreement and Bias Analysis

To understand how different architectures interpret the highly specialized lexicon of cryptocurrency news, we conducted a pairwise comparison using Cohen's Kappa (κ) as seen in Figure 2

The analysis of the inter-model agreement using Cohen's Kappa shown in Figure 2, reveals a consistent pattern across the Blockchain, Ethereum, and Altcoin sentiment datasets. While the majority of model pairs exhibit near-random agreement ($\kappa \approx 0.00$); highlighting a substantial divergence in how these architectures interpret the same cryptocurrency news. A distinct exception is observed between FinBERT and QWEN. Across all three subjects, this specific pair consistently demonstrates a weak-to-moderate positive correlation, suggesting that despite their structural differences, they share a more aligned understanding of market context compared to the purely crypto-native or legacy models.

3.2 Market Predictive Efficacy

The ultimate utility of financial sentiment analysis lies in its predictive power. We evaluated the daily aggregated sentiment signal of each model against the binary daily price action of Bitcoin (BTC), Ethereum (ETH), and the DIME Altcoin ETF.

3.3 Discussion

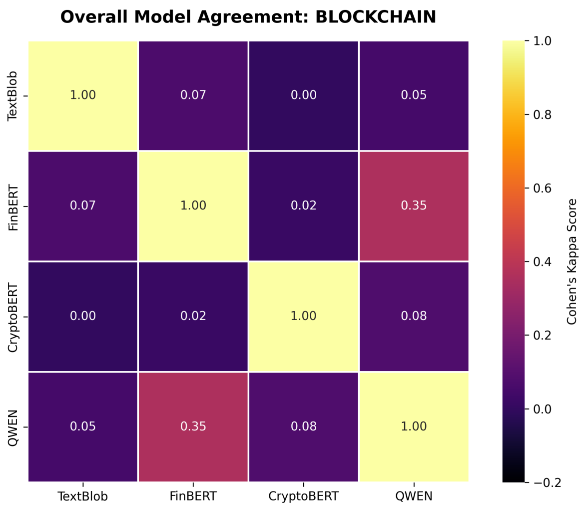
In evaluating the predictive efficacy of our sentiment pipelines, distinct performance tiers and asset-specific limitations emerged. On average, the domain-specific CryptoBERT exhibited the lowest predictive accuracy, underperforming even the traditional, lexicon-based TextBlob baseline. While FinBERT generated a mean predictive accuracy marginally above the 50% random baseline, its 95% confidence interval intersected with chance, rendering the performance edge statistically insignificant.

Notably, forecasting the Bitcoin (BTC) market proved universally challenging across all evaluated architectures. This uniform underperformance suggests that as a highly capitalized, macro-integrated asset, Bitcoin's price action is driven by a complex, multi-faceted blend of global economic factors that isolated news sentiment cannot independently capture.

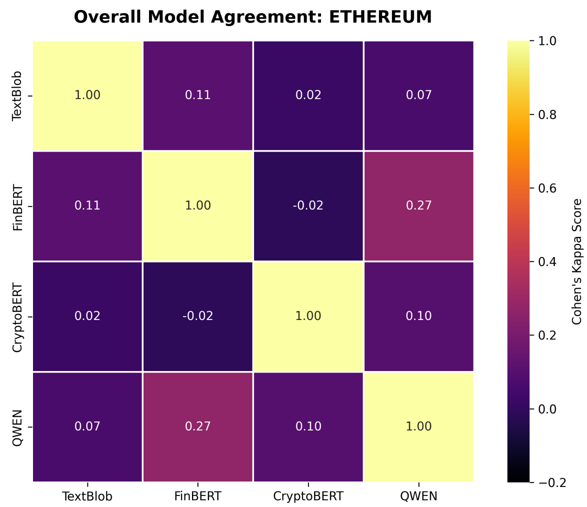
Conversely, the dense generative model, Qwen2.5-7B-Instruct, demonstrated robust out-of-the-box performance when applied to the Ethereum (ETH) and Altcoin (e.g., Dogecoin) datasets, where it showed an average accuracy of about 53%, while maintaining most of its 95% confidence interval, above the 50% random-guess threshold. These findings strongly indicate that while apex market-cap assets require multidimensional forecasting, advanced open-weight generative models like Qwen2.5 are highly capable of extracting actionable, predictive sentiment for secondary and alternative cryptocurrencies without the need for bespoke retraining.

4 Conclusion

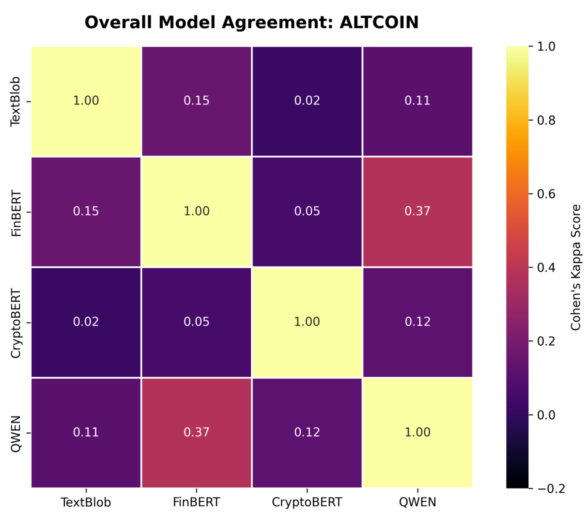
This study addressed the significant challenge of cryptocurrency market forecasting in an environment characterized by extreme volatility and historical data scarcity. Because the ephemeral nature of digital financial media makes training bespoke NLP models from scratch highly restrictive and computationally expensive, our primary objective was to determine whether existing, "open-box" architectures



(a) Blockchain (Tracking BTC)

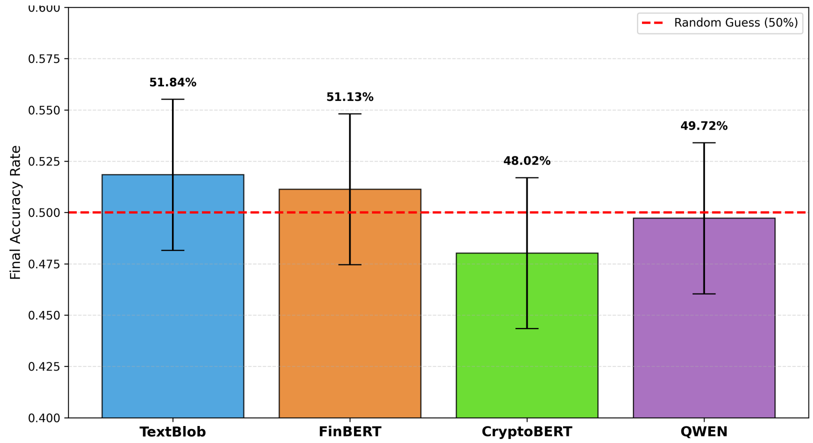


(b) Ethereum (Tracking ETH)

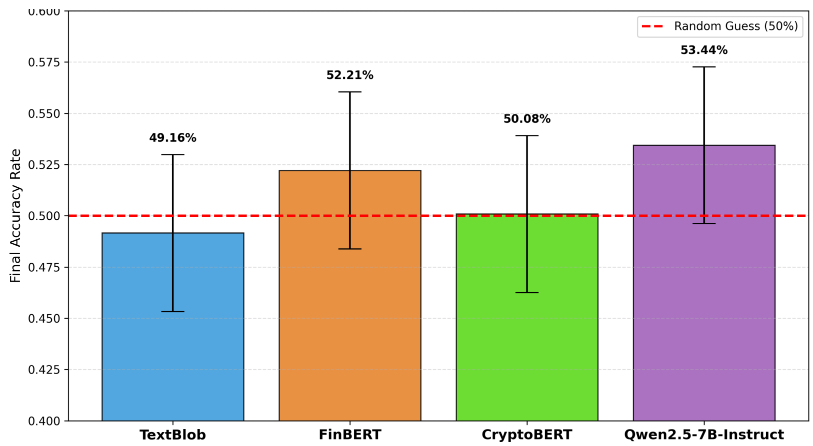


(c) Altcoin (Tracking DOGE)

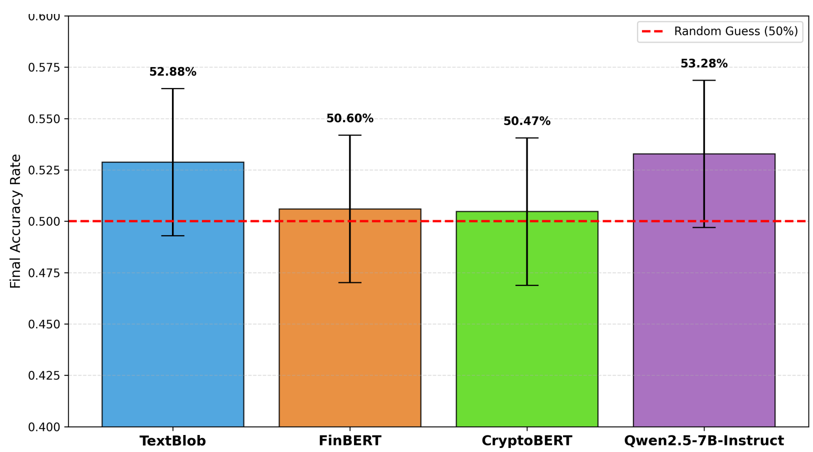
Figure 2: Inter-model agreement (Cohen's Kappa, κ) across three distinct cryptocurrency subjects. The heatmaps illustrate the pairwise correlation between the sentiment predictions of TextBlob, FinBERT, CryptoBERT, and QWEN. To evaluate asset-specific predictive performance, the news data was filtered by subject: Blockchain news was isolated to track Bitcoin (BTC), Ethereum news for Ether (ETH), and Altcoin news for Dogecoin (DOGE). Diagonal values represent perfect self-agreement ($\kappa = 1.00$). Off-diagonal values reveal significant divergences in model behavior; notably, all models exhibit near-random agreement ($\kappa \approx 0.00$) with TextBlob and CryptoBERT, while FinBERT and QWEN demonstrate a moderate positive correlation.



(a) Blockchain Sentiment vs. BTC Prices



(b) Ethereum Sentiment vs. ETH Prices



(c) Altcoin Sentiment vs. DOGE Prices

Figure 3: Directional predictive accuracy of the four sentiment models across three distinct cryptocurrency markets. Bar heights represent the final aggregate accuracy of each model predicting the $T + 1$ daily market direction, evaluated against a 50% random-chance baseline (red dashed line). Error bars denote the 95% confidence intervals ($\pm 1.96 \times SE$) for each prediction set.

could be effectively deployed out-of-the-box. By evaluating a diverse pipeline of domain-specific encoders, generative decoders, and baseline models on a 2021–2023 news dataset, we tested their capacity to extract actionable market signals without requiring ground-up retraining.

Our inter-model comparison revealed distinct behavioral differences among the architectures when exposed to the hyper-specific crypto lexicon. Notably, the Cohen's Kappa analysis demonstrated near-random agreement ($\kappa \approx 0.00$) among the majority of model pairs, highlighting a substantial divergence in how legacy and purely crypto-native models interpret the same text. However, a distinct exception emerged between FinBERT and QWEN. Across all three subjects, this pair consistently exhibited a weak-to-moderate positive correlation (reaching $\kappa = 0.37$ in the Altcoin dataset), suggesting that despite their structural differences, both models share a more aligned understanding of underlying financial risk and market context.

When these aggregated sentiment scores were benchmarked against actual price action, predictive efficacy varied significantly by asset class. While predicting Bitcoin's price movement proved highly resistant to isolated sentiment forecasting across all architectures, Qwen2.5-7B-Instruct emerged as a highly effective out-of-the-box forecaster for Ethereum and Altcoin markets. Furthermore, while the lower tails of the 95% confidence intervals for top-performing models occasionally dipped below the coin-flip threshold due to sample variance, the clear majority of their probability distributions remained situated above the 50% baseline. Conversely, purely crypto-native architectures like CryptoBERT demonstrated the weakest overall performance, acting with a heavy bias that failed to extract a genuine directional edge.

Ultimately, this research confirms that specific open-weight generative models, particularly parameter-efficient architectures like Qwen2.5-7B-Instruct, offer a viable, computationally accessible solution for crypto-economic sentiment analysis in alternative coin markets, successfully bypassing the bottleneck of historical data scarcity and bespoke retraining. Future research should explore the real-time streaming integration of these models into live algorithmic trading pipelines.

References

[1] İlayda İSABETLİ FİDAN and Tuğba GÜZ. The characteristics of cryptocurrency market volatility: Empirical study for five cryptocurrency. *Alphanumeric Journal*, 10(2):69–84, December 2022. ISSN 2148-2225. doi: 10.17093/alphanumeric.941529. URL http://dx.doi.org/10.17093/alphanumeric.941529.

[2] Zuzana Janková. Critical review of text mining and sentiment analysis for stock market prediction. *Journal of Business Economics and Management*, 24(1):177–198, April 2023. ISSN 2029-4433. doi: 10.3846/jbem.2023.18805. URL http://dx.doi.org/10.3846/jbem.2023.18805.

[3] Kelly Ann Coulter. The impact of news media on bitcoin prices: modelling data driven discourses in the crypto-economy with natural language processing. *Royal Society Open Science*, 9(4), April 2022. ISSN 2054-5703. doi: 10.1098/rsos.220276. URL http://dx.doi.org/10.1098/rsos.220276.

[4] Raphael Auer, Giulio Cornelli, Sebastian Doerr, Jon Frost, Leonardo Gambacorta, Jon Frost, and Leonardo Gambacorta. Crypto trading and bitcoin prices: Evidence from a new database of retail adoption. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4357559. URL http://dx.doi.org/10.2139/ssrn.4357559.

[5] Mengyu Wang, Shay B Cohen, and Tiejun Ma. Modeling news interactions and influence for financial market prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 3302–3314. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.189. URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.189.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

[7] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019. URL https://arxiv.org/abs/1908.10063.

[8] Prosus AI. Finbert: Financial sentiment analysis with pre-trained language models. https://huggingface.co/ProsusAI/finbert, 2019. Accessed via Hugging Face model hub.

[9] Mikolaj Kulakowski and Flavius Frasincar. Sentiment classification of cryptocurrency-related social media posts. *IEEE Intelligent Systems*, 38(4):5–9, 2023. ISSN 1941-1294. doi: 10.1109/mis.2023.3283170. URL http://dx.doi.org/10.1109/MIS.2023.3283170.

[10] ElKulako. Cryptobert. https://huggingface.co/ElKulako/cryptobert, 2024. Accessed via Hugging Face model hub.

[11] Qwen Team. Qwen2.5-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct, 2024. Accessed via Hugging Face model hub.

[12] Steven Loria. Textblob: Simplified text processing. https://textblob.readthedocs.io/en/dev/, 2018. Accessed for rule-based sentiment baseline.

[13] oliviervha. Crypto news dataset. https://www.kaggle.com/datasets/oliviervha/crypto-news, 2023.

[14] Ran Aroussi. yfinance: Yahoo! finance market data downloader. https://github.com/ranaroussi/yfinance, 2020.

[15] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[16] Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G. Constantinides, and Danilo Mandic. Finllama: Financial sentiment classification for algorithmic trading applications, 2024. URL https://arxiv.org/abs/2403.12285.