Predicting Cryptocurrency Price Trends Using Aspect-Based Sentiment Analysis and Economic Data

Kia Jahanbin[1], Mohammad Ali Zare Chahooki[2*], Hossein Fani[3], Abolfazl Razzaghi Yamchi[4†]

[1,2]Department of Computer Engineering, Yazd University, Yazd, Iran.
[3]School of Computer Science, Faculty of Science, University of Windsor, Windsor, Ontario, Canada.
[4]Department of Computer Science and Engineering, Shiraz University, Shiraz, Fars, Iran.

*Corresponding author(s). E-mail(s): chahooki@yazd.ac.ir;
Contributing authors: kia.jahanbin@stu.yazd.ac.ir;
hossein.fani@uwindsor.ca; razzaghi.abolfazl@gmail.com;
†These authors contributed equally to this work.

Abstract

The proliferation of blockchain technology has heightened interest in cryptocurrencies, which offer decentralized transaction capabilities independent of governmental institutions. Nevertheless, cryptocurrency markets are characterized by high capital intensity and pronounced price volatility, rendering them both lucrative and risky. Consequently, predicting price trends in these markets has emerged as a prominent yet challenging research area. This paper proposes an adaptive predictive model for forecasting price trends of prominent cryptocurrencies by integrating aspect-based sentiment analysis derived from Telegram channels with key economic indicators. The model employs a feature engineering process that utilizes Spearman's rank correlation coefficient to identify the most relevant predictive features for each cryptocurrency. Subsequently, the selected features are balanced using the Synthetic Minority Over-sampling Technique (SMOTE) and incorporated into a finely-tuned XGBoost classifier, which is specifically optimized for each digital asset. Notably, the prediction horizon is treated as an additional hyperparameter, enabling the model to identify and report the optimal time frame for trend prediction based on peak performance metrics. Experimental results demonstrate that the proposed model achieves an average accuracy of 70% and an F1-Score of 73%, outperforming existing benchmarks by a margin of 2–10% in accuracy for the majority of cryptocurrencies evaluated.

Keywords: Aspect-Based Sentiment Analysis (ABSA), Cryptocurrency, Feature Engineering, Hyperparameter Optimization, Price Trend Forecasting, XGBoost

1 Introduction

Nowadays, cryptocurrencies have become a popular topic in economics and investment. Their popularity stems from advantages such as fast transaction speed, lack of dependence on governments, a 24/7 market, and highly secure transfer and transaction tracking. However, due to their high price volatility, investing in leading cryptocurrencies such as Bitcoin and Ethereum and their considerable capital does not generate much profit for the average investor. Therefore, both risk-tolerant and many ordinary investors are interested in investing in emerging tokens, which exhibit high price volatility. The price trend of emerging tokens often depends on the state of the market, and their level of acceptance by investors [1].

While the price trends of leading cryptocurrencies like Bitcoin are effectively analyzed through technical rules and macroeconomic parameters, emerging tokens are far more reactive to market psychology?namely, greed and fear.This reliance on sentiment makes the opinions of expert influencers a potential key driver of market acceptance. Yet, most major influencers avoid commenting on emerging tokens due to their highly volatile nature. Consequently, Twitter and Reddit offer limited insight into these trends. For this crucial information, investors turn to large Telegram groups, which have become essential sources for opinions and analysis on all growing—or currently trending—cryptocurrencies. These analyses are a primary support for the decision-making of ordinary investors.

Traditional price prediction models, which rely predominantly on historical prices, are often ineffective for cryptocurrencies due to the market's high volatility and lack of seasonality [2]. To address this limitation, our research proposes a novel approach that integrates three distinct categories of data: historical prices, technical analysis indicators, and–crucially–market behavior. We quantify market behavior by performing a weighted sentiment analysis on comments from Telegram groups, thereby capturing the 'fear and greed' dynamics that significantly influence trending cryptocurrencies.

We begin by extracting the daily trending tokens using the CoinGecko API. For these tokens, we then scrape comments from major cryptocurrency analysis Telegram channels; the selection of these channels is informed by reports from authoritative sources like Techopedia and CoinGape. Following data cleaning, we perform sentiment analysis on the comments using Aspect-Based Sentiment Analysis (ABSA) via the HDRB model [3]. This hybrid model, which integrates RoBERTa and a BiGRU architecture, allows us to infer market behavior for each cryptocurrency. A key feature of the model is its ability to weight influential opinions more heavily, assigning greater significance to the polarity of comments that receive stronger positive reception from users.

Technical indicators were calculated using the `ta` library. Daily historical price data was retrieved via the Yahoo Finance API. Rather than predicting exact prices, this research focuses on forecasting the directional price trend (up or down) for each cryptocurrency over a horizon of 1 to 10 days. This forecast horizon is treated as a model hyperparameter. The optimal prediction time frame for each asset is determined empirically based on peak model accuracy and reported alongside the trend forecast. For instance, the optimal horizon for the ARB token is three days, whereas for the DOT token, it is four days. Our approach employs a fine-tuned XGBoost model for trend prediction. For each cryptocurrency, the optimal set of XGBoost hyperparameters is identified using a Grid Search technique, and the model is subsequently trained using this configuration.

Furthermore, feature engineering is employed to identify the most relevant features influencing the price trend of each cryptocurrency. Specifically, we first select the features most pertinent to the target variable using Spearman's rank correlation coefficient. This selected feature set is then used to train the XGBoost model. Subsequently, we validate the importance of these features using SHAP (Shapley Additive Explanations) [4], a method rooted in cooperative game theory that quantifies the contribution of each feature to the model's predictions, thereby providing a deeper analysis of the dataset and the underlying problem.

The significant contributions of this research are as follows:

1. **Targeted Investment Insight:** It identifies investment opportunities in trending, emerging cryptocurrencies, specifically catering to the profile of high-risk investors.
2. **Interpretable Feature Selection:** It introduces a robust feature engineering pipeline that combines Spearman?s correlation for initial selection with SHAP analysis for model-based validation, precisely identifying the most influential features for each cryptocurrency's price trend.
3. **Adaptive Prediction Horizon:** It frames the prediction time frame as a tunable hyperparameter, dynamically identifying the optimal forecast horizon (1-10 days) for each cryptocurrency to maximize trend prediction accuracy.
4. **Specialized Model Architecture:** It develops a customized forecasting approach that fine-tunes a separate XGBoost model for each cryptocurrency and integrates weighted aspect-based sentiment analysis from Telegram to accurately detect market behavior.

This paper elaborates on the aforementioned contributions through the following structure. First, Section 2 reviews the literature on cryptocurrency trend and price prediction. The proposed model, Adaptive Trend Crypto Prediction with HDRB and XGBoost (ATCPHX), is then detailed in Section 3. The experimental setup and results are presented and discussed in Section 4. Finally, Section 5 concludes the paper and suggests directions for future research.

2 Related Works

Following the taxonomy proposed by Poongodi et al. [5], research on cryptocurrency market prediction can be categorized into three groups based on the type of data used: (1) price data only, (2) price and social data, and (3) price, social, blockchain, and economic data. This study falls into the second category, focusing specifically on Telegram social network data and technical indicators. Accordingly, this section first provides a brief overview of studies utilizing only price data. It then reviews related work that incorporates social network data and other data types (i.e., categories 2 and 3) to establish the background and position of our research.

2.1 Use of Price Data in Cryptocurrency Price Forecasting

Using only historical price data is the most common and foundational method for predicting financial markets, including cryptocurrencies. However, a key limitation of price-only models is their inability to capture market sentiment. To address this, researchers have incorporated social network data, hypothesizing that investor comments and discussions significantly influence price trends [5]. The following section will first briefly review seminal articles that rely solely on previous price data before examining these more advanced approaches.

The research by Chen et al. [6] examines Bitcoin price forecasting through machine learning, with a specific focus on the influence of temporal resolution. The authors demonstrate a notable divergence in model performance: for daily price predictions with high-dimensional features, statistical methods achieved higher accuracy (66%) than complex ML algorithms. In contrast, for 5-minute high-frequency forecasting, advanced techniques–including Random Forest, XGBoost, Quadratic Discriminant Analysis, SVM, and LSTM–proved superior, reaching 67.2% accuracy. This study highlights the critical role of data sampling frequency and feature space dimensionality in determining the optimal modeling approach for financial forecasting.

Poongodi et al. [7] utilized Bitcoin's network data and examined various machine learning algorithms for prediction, framing their study within the context of Bitcoin's growing digital prominence and volatility. Their research focuses on the growing trend of users joining major Bitcoin mining pools to mitigate the high variance of solo mining, thereby enhancing their chances of earning rewards through block generation. Within this context, the authors applied the ARIMA model to analyze and predict Bitcoin's price movements, offering insights into its applicability for financial time-series analysis in cryptocurrency markets.

Patel et al. [8] focused on cryptocurrency price prediction using advanced machine learning and deep learning techniques, including Gated Recurrent Units (GRU), Neural Networks (NN), and Long Short-Term Memory (LSTM). They proposed a hybrid model combining LSTM and GRU, specifically applied to Litecoin and Monero. Their findings demonstrated high predictive accuracy, suggesting the model's potential applicability to other cryptocurrencies and offering valuable insights for the field of cryptocurrency forecasting.

Vo et al. [9] introduced a method for predicting cryptocurrency prices, with a primary focus on Ethereum, by integrating news sentiment and historical price data.

Table 1: Literature review of cryptocurrency price prediction based on price data.

Study	Key Contribution	Limitations
Chen et al. (2020) [6]	Explores ML for Bitcoin price prediction using high-dimensional features (network, trading, market attention) across daily and high-frequency data.	Focused solely on Bitcoin. Limited exploration of modeling techniques for different data structures.
Poongodi et al. (2020) [7]	Analyzes Bitcoin's volatility and its digital transformation. Applies the ARIMA model for price forecasting.	Exclusively focuses on Bitcoin. Relies only on ARIMA without comparing other methods.
Patel et al. (2020) [8]	Investigates price prediction for Litecoin and Monero using LSTM and GRU models alongside blockchain technology.	Limited to Litecoin and Monero. Lacks methodological diversity beyond cryptographic algorithms.
Vo et al. (2019) [9]	Proposes using NLP for sentiment analysis to predict cryptocurrency prices.	Utilizes traditional price prediction models that rely only on historical data.

Their model generates trading signals–buy, sell, or hold–by employing natural language processing (NLP) algorithms to perform sentiment analysis on news data. The study highlights the significant role of sentiment analysis in cryptocurrency price prediction, demonstrating that combining sentiment insights with historical data effectively forecasts price direction.

A comparative summary of contributions and limitations in cryptocurrency price forecasting research from the discussed studies is presented in Table 1, offering a concise overview of methodological strengths and weaknesses across the literature.

2.2 Using Price Data and Social Networks in Cryptocurrency Price Prediction

Social network data has emerged as a significant factor in determining market behavior, particularly for cryptocurrency markets where sentiment plays a crucial role. While existing research [1, 5] has predominantly utilized Twitter data to gauge market sentiment through cryptocurrency-specific hashtags, this approach captures both expert and non-expert opinions, potentially diluting analytical quality. Although tweets containing these hashtags are cryptocurrency-related, they often include content from casual observers rather than knowledgeable analysts. An alternative approach focuses on influencer commentary, which can be more effective for determining price trends; however, influencers typically concentrate on established cryptocurrencies like Bitcoin, Ethereum, and Ripple while avoiding emerging tokens due to their volatility and authenticity concerns, creating a significant gap in social media-based prediction for these newer assets

Davchev et al. [10] utilized an existing Kaggle tweet dataset alongside historical Bitcoin price data for price prediction. In their methodology, Bitcoin related tweets polarity was determined using a pre-trained RoBERTa neural network for preprocessing. The sentiment values derived from tweet polarity were then integrated with past Bitcoin price data and used as input to FbProphet and XGBoost regressors to forecast future prices.

Abraham et al. [11] employed a hybrid LSTM+Gated Recurrent Unit (GRU) model to predict prices for Monero and Litecoin. Their model incorporated multiple data sources as inputs, including stock market values, trading volume, historical prices, Google Trends search volume, and Twitter commentary.

Li et al. [12] employed a long short-term memory (LSTM) model with multiple input features–including blockchain metrics, Twitter data, and Google Trends data–to predict Bitcoin's price volatility over a 30-day horizon. While this approach incorporates diverse data sources, a significant research gap remains in the lack of model accuracy adjustment specifically tailored for sentiment analysis, potentially limiting the predictive precision of social media-derived features.

Huang et al. [13] investigated cryptocurrency price prediction by analyzing sentiment in social media posts, with a focus on addressing highly volatile price movements. Their work extends existing research–which predominantly focuses on English-language sentiment analysis–by developing a novel method to evaluate sentiments in Chinese social media posts from Sina-Weibo. The methodology involves collecting Weibo posts, constructing a cryptocurrency-specific sentiment dictionary, and deploying an LSTM-based recurrent neural network integrated with historical price data. The study aims to accurately forecast future cryptocurrency price trends. Experimental results demonstrate that this approach outperforms state-of-the-art auto-regressive models, achieving improvements of 18.5% in precision and 15.4% in recall.

Wołk [14] utilized Twitter and Google Trends in their study to forecast short-term cryptocurrency prices, acknowledging the influence of these platforms on investment decisions. This paper adopts a unique multi-modal approach to examine the impact of social media on cryptocurrency prices. The findings underscore the substantial effect of psychological and behavioral attitudes on the speculative nature of cryptocurrency prices, demonstrating a significant correlation between public sentiment and market movements.

Amirshahi et al. [15] investigated cryptocurrency price prediction, addressing a topic of growing interest among investors, researchers, and practitioners. The authors proposed a hybrid model combining long short-term memory (LSTM), convolutional neural network (CNN), and attention mechanisms to forecast daily cryptocurrency closing prices. Their analysis revealed that longer input sequences (14- and 21-day periods) generally yielded more accurate predictions than shorter 7-day histories, though minimal difference was observed between the 14- and 21-day windows. Furthermore, the study demonstrated that sentiment data significantly enhanced prediction accuracy for over 70% of the examined cryptocurrencies.

Critien et al. [16] advanced Bitcoin price prediction methodologies by simultaneously determining both the direction and magnitude of price movements. Their approach integrates sentiment analysis and tweet volume metrics, employing both recurrent and convolutional neural networks. The authors introduced a multi-class classification framework specifically designed to predict price fluctuation magnitudes. A significant contribution of their research is the identification of optimal time intervals during which sentiment data most effectively signals impending price changes. The proposed model achieves approximately 63% accuracy in predicting both directional trends and magnitude of Bitcoin price changes.

Parekh et al. [17] introduce DL-GuesS, a robust hybrid framework for cryptocurrency price prediction that accounts for inter-dependencies among cryptocurrencies and market sentiments derived from tweets and social media. Focusing specifically on Dash and Bitcoin Cash, their model incorporates historical price data and relevant tweets from Dash, Litecoin, and Bitcoin, utilizing multiple loss functions for validation. The framework addresses the challenge of predicting highly volatile cryptocurrency prices while emphasizing the interconnected nature of cryptocurrency markets.

Low et al. [18] present the Deep Learning Cryptocurrency Forecasting with Sentiment (DLCFS) system, a novel framework that integrates deep learning architectures with sentiment analysis for cryptocurrency price prediction. Designed to address the high volatility of cryptocurrencies such as Bitcoin, Ethereum, and Litecoin, DLCFS analyzes historical price data and Reddit submissions to generate forecasts. The system demonstrates superior performance compared to traditional machine learning models, achieving high correlation coefficients that validate its effectiveness. This work highlights the critical importance of incorporating market sentiments and cross-asset interdependencies in cryptocurrency forecasting models.

Roy et al. [19] proposed a deep learning-based framework for Bitcoin price prediction, addressing the critical need for accurate modeling given the cryptocurrency's inherent non-linearity and extreme volatility. The authors developed a Long Short-Term Memory (LSTM) model that outperforms comparable deep learning approaches by achieving minimal loss metrics. Their results demonstrate the model's efficacy in forecasting future cryptocurrency prices, offering valuable insights for global investors and industries seeking informed investment strategies. This research underscores the significance of advanced predictive modeling in navigating the complex and uncertain dynamics of cryptocurrency markets.

A summary of the key innovations and research gaps from the discussed studies is presented in Table 2. The table highlights methodological contributions and limitations in social-media-driven cryptocurrency price prediction, contextualizing advances and unresolved challenges in the field. Building upon the critical review of existing studies in social-media-driven cryptocurrency price prediction, the following key research gaps have been identified in the current literature:

1. A narrow focus on major cryptocurrencies (e.g., Bitcoin, Ethereum), which fails to address the opportunities in trending or emerging tokens that may appeal to different investor profiles.
2. Reliance on lexicon-based sentiment analysis tools (e.g., VADER) that are not fine-tuned for the cryptographic domain, leading to suboptimal text polarity extraction.
3. A lack of aspect-based sentiment analysis, which prevents the extraction of nuanced opinions on specific features of cryptocurrencies.
4. Neglect of user expertise and influence metrics when extracting opinions from social networks, treating all comments as equally informative.

Table 2: Literature review of cryptocurrency price prediction based on social media, price, and financial data.

Study	Key Contribution	Limitations
Wołk (2020) [14]	Uses Twitter and Google Trends for short-term prediction with a multi-modal approach to quantify social media impact.	May over-rely on social sentiment, potentially overlooking other fundamental market drivers.
Li et al. (2019) [12]	Combines an LSTM model with the Black-Scholes formula for Bitcoin option pricing, using blockchain statistics and social trends.	Reliance on specific models (LSTM) and data may not fully capture broader market complexities.
Davchev et al. (2020) [10]	Predicts Bitcoin price using Twitter data and historical prices via NLP models and transfer learning for sentiment analysis.	Focused primarily on Bitcoin; methodology may not generalize well to other cryptocurrencies.
Critien et al. (2022) [16]	Predicts Bitcoin's price direction and magnitude using neural networks on Twitter sentiment and volume data.	Heavy reliance on Twitter data may miss other crucial factors affecting price fluctuations.
Parekh et al. (2022) [17]	Integrates deep learning with sentiment analysis for enhanced cryptocurrency price prediction.	Utilizes traditional models that rely predominantly on historical price data.
Abraham et al. (2018) [11]	Predicts Bitcoin and Ethereum prices using Twitter and Google Trends data, identifying tweet volume as a key predictor.	Over-reliance on tweet volume may overlook other significant influencing factors.
Low et al. (2023) [18]	Proposes DLCFS, a deep learning system using Reddit submissions and market interdependencies for prediction.	Focus on specific cryptocurrencies and platforms may limit general applicability.
Roy et al. (2023) [19]	Proposes a method for price prediction using news and historical data, applying NLP for sentiment analysis on Ethereum.	Reliance on news data may not fully capture overall market sentiment.
Amirshahi et al. (2023) [15]	Proposes a hybrid LSTM-CNN model with attention mechanisms for daily price prediction across 27 cryptocurrencies.	High model complexity may hinder practical deployment in real-world scenarios.
Huang et al. (2021) [13]	Describes a prediction method using news and historical data, focusing on Ethereum with NLP-based sentiment analysis.	Dependence on news data may not be fully representative of broader market sentiment.

3 Methodology

This section details the Adaptive Trend Crypto Prediction with HDRB and XGBoost (ATCPHX) model, a framework based on transfer learning and a suite of fine-tuned machine learning algorithms. The ATCPHX pipeline initiates by extracting trending tokens from the market using the CoinGecko API. In this research, the term trending cryptocurrencies specifically denotes tokens identified by this API, which highlights assets gaining significant market attention based on a multifaceted ranking system. This system synthesizes (i) user search interest across major exchanges and platforms, (ii) short-term transaction and trading volume, and (iii) social momentum derived from community activity. Consequently, our definition of "trending" extends beyond mere price volatility or raw trading volume, encompassing instead a composite view of a token's economic engagement and its social visibility within the market.

Subsequently, analysis content related to these tokens is scraped from prominent Telegram channels. Additionally, key economic indicators–including technical analysis metrics and price volatility–are retrieved using the CFGI.io API (a detailed explanation of these indicators is provided in Section 4). Historical price data for each token is also obtained through the Yahoo Finance API.

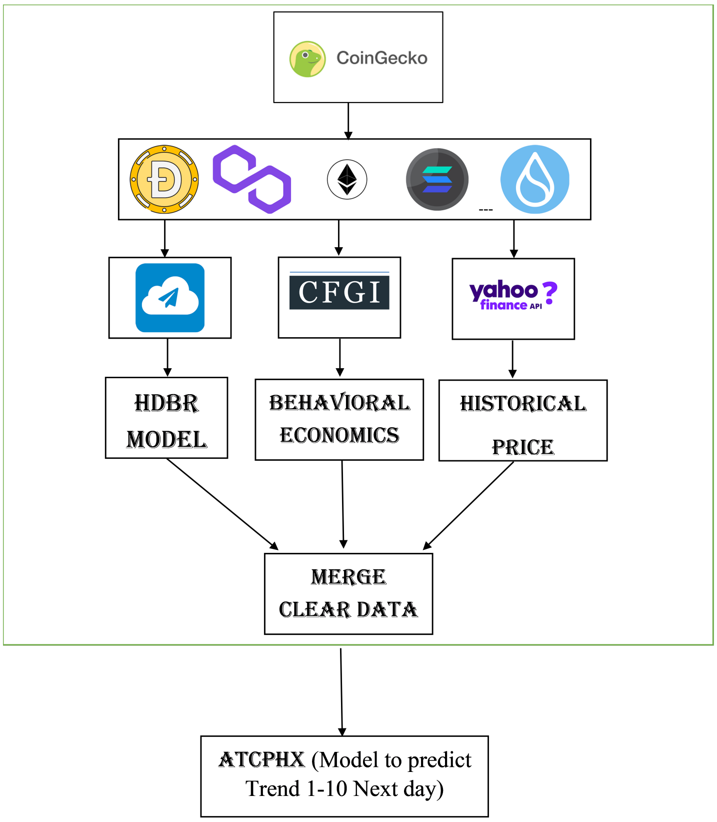

Fig. 1: Main framework

The complete data acquisition and integration process is illustrated in Figure 1, demonstrating the flow from multi-source data collection to model input preparation. The data processing pipeline, as illustrated in Figure 1, begins with the extraction of trending tokens via the CoinGecko API. This includes both established main coins and emerging tokens exhibiting high popularity across exchanges. Following cryptocurrency identification, relevant economic indicators and market sentiment data are collected for trend prediction. Financial metrics?including volatility, impulse, technical indicators, and whale activity?are obtained from the CFGI API, while historical OHLCV price data is sourced from Yahoo Finance.

Market sentiment is evaluated through content extracted from 15 prominent Telegram channels, with aspect-based sentiment analysis performed using the HDRB model [3]. This model employs transfer deep learning to extract textual aspects and determine their respective polarities. All collected data are merged into a unified dataset, and a target column is incorporated to indicate daily price trends. Specifically, the trend label is assigned a value of 1 if the closing price on day $t + 1$ exceeds that on day t, and 0 otherwise.

The merged dataset is then used within the ATCPHX framework to determine price trends over horizons of 1 to 10 days. Feature engineering is first conducted per cryptocurrency using Spearman's correlation coefficient, which is selected due to its non-parametric nature and suitability for features with non-normal distributions. Subsequently, the selected features are used to train an XGBoost model, with hyperparameters fine-tuned via GridSearch for each cryptocurrency. Given the manageable dataset size and computational efficiency of XGBoost, this process remains feasible without excessive overhead. The fine-tuned models are saved for future inference, eliminating the need for repeated retraining. The following subsections detail the HDRB sentiment analysis methodology and the feature selection process using Spearman's correlation.

3.1 Aspect-Based Sentiment Analysis Using the HDRB Model

Aspect-based sentiment analysis (ABSA) aims to computationally identify and categorize opinions within text by focusing on specific aspects of entities being discussed. Formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ represent a corpus of documents or comments, where each document d_i consists of a sequence of sentences $\mathcal{S} = \{s_1, s_1, \ldots, s_m\}$. The goal of ABSA is to learn a mapping $f : \mathcal{S} \times \mathcal{A} \to \mathcal{O}$ where $\mathcal{A} = \{a_1, a_1, \ldots, a_k\}$ denotes the set of aspects related to the target entity, and $\mathcal{O} = \{o_1, o_2, \ldots, o_l\}$ represents possible sentiment polarities (e.g., positive, negative, neutral) associated with each aspect a_j a given sentence s_i.

The comments extracted from Telegram channels undergo two sequential preprocessing phases: general and exclusive, applied after data collection. The general preprocessing phase consists of the following operations:

1. **Equalization:** All characters in the text are converted to lowercase to ensure uniformity. Formally, for a text T comprising words $w_1, w_2, \ldots, w_n$, lowercasing transforms each word w_i to its lowercase form $\tilde{w}_i$, resulting in a normalized text $\tilde{T} = \tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n$.
2. **Stop Word Removal:** Commonly used words that contribute minimal semantic meaning (e.g., "the", "s", "at") are filtered out. Given a set of stop words S and a text corpus T, the process yields a new corpus $\tilde{T} = T \setminus S$.
3. **Punctuation Removal:** All punctuation marks (e.g., commas, periods, exclamation marks) are deleted. For a text T and a set of punctuation symbols P, the resulting text is $\tilde{T} = T \setminus P$.
4. **Lemmatization:** Each word is reduced to its base dictionary form (lemma) using morphological analysis. Unlike stemming, lemmatization considers the context and morphological analysis of words, thus ensuring that the root word (lemma) belongs to the language. For a word w, its lemma $L(w)$ is derived considering linguistic context and morphological rules.
5. **Stemming:** Words are reduced to their root form by removing affixes (e.g., suffixes, prefixes) through heuristic methods. Unlike lemmatization, it ignores linguistic context and may produce non-valid words. The stem $S(w)$ of a word w may not always be a valid word, as this process does not consider linguistic context.

6. **Content Filtering and Relevance Scoring:** To mitigate noise from bot-generated or irrelevant content, we applied a multi-step filtering process:
   - Removal of promotional messages, advertisements, and duplicate posts containing only URLs.
   - Application of user-based heuristics: comments from accounts with high posting rates but low engagement (e.g., low like/view ratios) were flagged as bot-like and excluded.
   - Retention of comments with at least three words and those containing cryptocurrency-related keywords.
   - Assignment of an importance coefficient to each comment based on user engagement (likes, views) and channel popularity, thereby reducing the weight of low-engagement bot-like comments.

In the dedicated preprocessing phase, each Telegram comment was assigned an importance coefficient (ranging from 0 to 1) based on its number of views, likes, and the size of the channel's user base. This coefficient reflects the relative influence or relevance of the comment. The corpus was tokenized using RoBERTa tokenization prior to this weighting step.

In defining the importance coefficient for a Telegram comment based on the number of views (V), likes (L), and the number of channel users (U), we can use a mathematical equation that combines these metrics to produce a value between 0 and 1, representing the importance of each comment. Let V, L, and U represent the number of views, likes, and channel users, respectively. A method to calculate the importance coefficient is to use a weighted average of these metrics, normalized to the range [0, 1], as shown in Equation (1):

$$\text{Importance Coefficient} = \frac{w_v \cdot V + w_l \cdot L + w_u \cdot U}{w_v \cdot V_{\max} + w_l \cdot L_{\max} + w_u \cdot U_{\max}} \tag{1}$$

where:

- V, L, U: Raw counts of views, likes, and channel users for the comment.
- $V_{\max}, L_{\max}, U_{\max}$: Maximum observed values of views, likes, and channel users in the dataset.
- w_v, w_l, w_u: Non-negative weights assigned to views, likes, and channel users, respectively, reflecting their relative importance.

The weights w_v, w_l, and w_u should be assigned based on the relative importance of each metric in determining a comment's significance. For illustrative purposes, we assign the following values reflecting our domain-specific assumptions: $w_v = 0.5$, $w_l = 0.3$, and $w_u = 0.2$. Here, $V_{\max}$, $L_{\max}$, and $U_{\max}$ represent the maximum possible values of views, likes, and users (either platform limits or observed maxima in the dataset) used for normalization. This formulation ensures that the importance coefficient is bounded in [0,1], remains interpretable, and allows flexibility in emphasizing

Table 3: Configuration setup for the HDRB model

Parameter	**Value**
Tokenizer and pre-trained model	Twitter-roberta-base
Dropout	One layer, 0.5 rate
BiGRU	128 units
Attention	Self-Attention
Activations	ReLU, Softmax
Optimizer	Adam(`learning_rate=1e-5`)
Loss	Categorical cross-entropy
Epochs	50
EarlyStopping	monitor='val_loss', patience=5

different engagement signals. The weights can be further refined using domain expertise or empirical optimization (e.g., grid search or regression against market response metrics).

Equation (1) quantifies the importance of a comment by integrating user engagement (views and likes) with community reach (channel size). This approach assigns higher importance coefficients to comments exhibiting greater interaction (e.g., views and likes) and those originating from larger channels, reflecting their broader potential impact. The method provides an objective, data-driven measure of relative comment significance, ensuring consistency and transparency in the weighting process. By relying on quantifiable metrics, it reduces subjective bias and supports reproducible analysis.

After preprocessing, the text is passed to the HDRB model for aspect-based sentiment analysis. The architecture of the HDRB model (Figure 2) constitutes a hybrid deep learning framework tailored for cryptocurrency-related ABSA, integrating the advantages of Concept-Latent Dirichlet Allocation (Concept-LDA) [20] for precise aspect extraction with a powerful deep learning backbone. The model features a pre-trained RoBERTa layer to capture rich contextual representations from the input text. The tokenized input sequence $X = \{x_1, x_2, \ldots, x_n\}$ s processed by RoBERTa to generate contextualized embeddings $H_{roberta} = RoBERTa(X)$.

The hyperparameters and configuration of the HDRB model [3] used for the ABSA task are detailed in Table 3. EarlyStopping is employed to monitor the training process and prevent overfitting[1].

Following the RoBERTa layer, the model utilizes a Bidirectional Gated Recurrent Unit (BiGRU) to capture sequential dependencies within the data. This is formally represented as $H_{bi\_gru} = BiGRU(H_{roberta})$. The final component of the HDRB architecture is an attention mechanism, which directs the model's focus to the most salient parts of the input text for sentiment analysis [3]. This ensures that influential tokens and phrases contribute more significantly to the final prediction.

3.2 ATCPHX Model

The ATCPHX model employs a feature engineering process where Spearman's rank correlation coefficient is used to measure the relationship between input features and

[1] For more reading, refer to Jahanbin et al. [3].

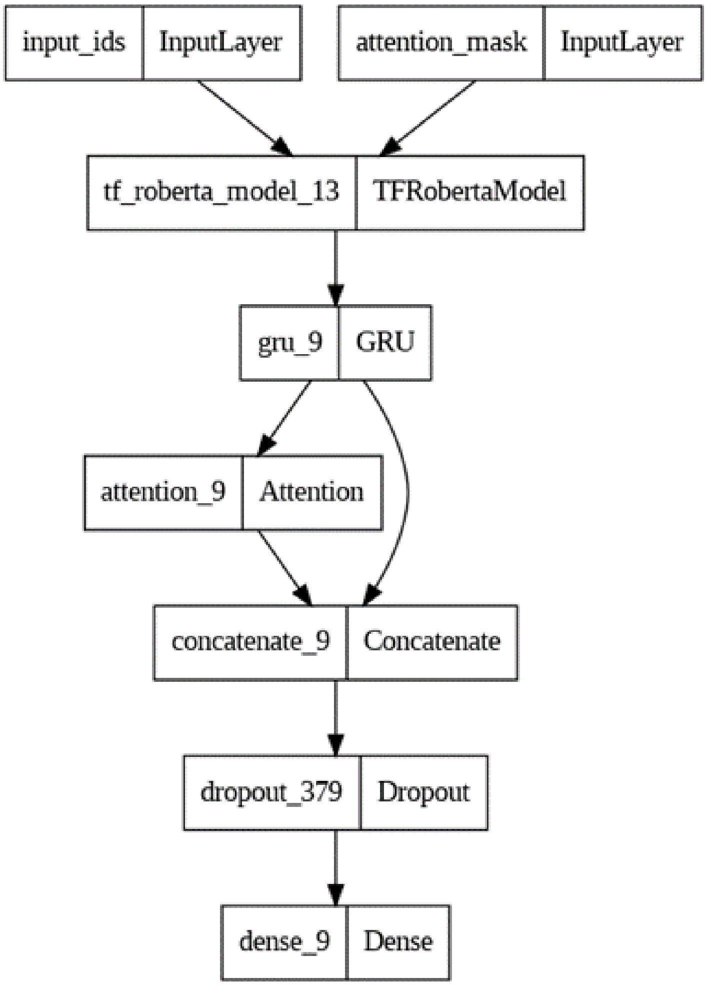

Fig. 2: Architecture of the HDRB model for aspect-based sentiment analysis of Telegram comments.

the target variable (currency trend). The input features include: 1) historical price data, 2) sentiment polarity from social networks, and 3) economic indicators such as technical scores. Any feature with a correlation coefficient below 0.4 is excluded from model training. This process is applied individually to each cryptocurrency, resulting in a unique feature set per asset. For instance, the "Close price" feature may be retained for cryptocurrency X but removed for cryptocurrency Y, depending on its correlation with the target trend variable. The feature selection procedure is outlined in Algorithm 1. The pseudocode for this process would involve iterating over each cryptocurrency in the dataset and, within each iteration, examining each feature across the specified categories. For each feature, the Spearman's rank correlation coefficient

Algorithm 1 Feature Selection using Spearman's Correlation

Require: Dataset D with features F, Target variable y, Threshold $\tau = 0.4$
Ensure: Filtered feature set F_{selected}

procedure FEATURESELECTION(D, F, y)
  $F_{\text{selected}} \leftarrow F$ ▷ Initialize with all features
  for each cryptocurrency $c \in D$ **do**
    for each feature $f \in F$ **do**
      $\rho \leftarrow \text{SpearmanCorrelation}(f, y_c)$ ▷ Calculate correlation
      if $\rho < \tau$ **then**
        $F_{\text{selected}} \leftarrow F_{\text{selected}} \setminus \{f\}$ ▷ Remove low-correlation feature
      end if
    end for
  end for
  return F_{selected}
end procedure

with the target column is calculated. If this correlation is found to be less than 40%, the feature is removed from the set considered for model training. After feature selection, the ATCPHX model is trained with the remaining features.

The feature selection process can be formally defined as follows: Let C be the set of all cryptocurrencies in the dataset, and F the set of features (e.g., past prices, social media sentiment polarity, and economic indicators). For each cryptocurrency $c \in C$ and each feature $f \in F$, the Spearman?s rank correlation coefficient $\rho(f, y_c)$ is computed, where y_c denotes the target variable (price trend) for cryptocurrency c. Features are selected based on the condition:

$$\rho(f, y_c) \geq 0.4$$

If $\rho(f, y_c) \leq 0.4$, the feature f is excluded from the training set for cryptocurrency c. This criterion is applied uniformly across all cryptocurrencies and features, ensuring only statistically relevant predictors are retained for model training.

The time complexity of the feature selection algorithm is $O(n \times m)$, where n is the number of cryptocurrencies and m is the number of features. This complexity arises from the nested iteration over all n cryptocurrencies and all m features, where each pair (c, f) requires computation of the Spearman's rank correlation coefficient. While the calculation of Spearman's ρ for a single feature-cryptocurrency pair has a cost of $O(T \log T)$ (where T is the number of observations per feature), the dominant term in the overall complexity remains $O(n \times m)$, assuming T is bounded and relatively small compared to n and m. This approach ensures scalability for practical datasets typical in cryptocurrency analysis.

After selecting the independent variables for each cryptocurrency, the Synthetic Minority Over-sampling Technique (SMOTE) [21] is applied to mitigate class imbalance in the dataset. SMOTE generates synthetic samples of the minority class by interpolating between existing instances in feature space. Formally, for a minority class sample x, SMOTE selects a random neighbor $\tilde{x}$ from its k-nearest neighbors and

Algorithm 2 Per-Asset XGBoost Tuning and Hyperparameter Optimization

procedure FINETUNEXGBOOST(*cryptocurrencies*)
  for each *crypto* in *cryptocurrencies* **do**
    $hyperparameter\_grid \leftarrow \{\text{max\_depth} : [3, 4, 5],$
      $\text{learning\_rate} : [0.005, 0.01, 0.02],$
      $\text{n\_estimators} : [250, 300, 350],$
      $\text{subsample} : [0.7, 0.8, 0.9],$
      $\text{min\_child\_weight} : [1, 3, 5],$
      $\text{gamma} : [0, 0.1, 0.2]\}$
    $model \leftarrow \text{XGBClassifier}()$
    $grid\_search \leftarrow \text{GridSearchCV}(model, hyperparameter\_grid)$
    $grid\_search.fit(X^{train}_{crypto}, y^{train}_{crypto})$
    $optimized\_model \leftarrow grid\_search.best\_estimator\_$
    $\text{SAVE}(optimized\_model, \text{path}/crypto\_model.pkl)$
  end for
end procedure

constructs a new sample x_{new} as follows:

$$x_{\text{new}} = x + \lambda \cdot (\tilde{x} - x) \tag{2}$$

where λ is a random number uniformly distributed in $[0, 1]$. This process repeats until the class distribution is balanced. By augmenting the minority class, SMOTE reduces bias toward the majority class during model training, thereby improving generalization performance on unseen data. The technique is implemented using the `imblearn.over_sampling` package in Python.

The balanced dataset is subsequently used as input to the XGBoost model, where hyperparameters are optimized for each cryptocurrency individually via GridSearch. Each fine-tuned model is persisted for future use, eliminating the need for retraining. The complete optimization procedure is formalized in Algorithm 2.

Let C represent the set of all cryptocurrencies. For each cryptocurrency $c \in C$, he XGBoost model is initialized and fine-tuned. The hyperparameter space H defined in Algorithm 2, is explored using GridSearchCV to identify the optimal configuration for each cryptocurrency-specific model. This tailored approach ensures that both feature selection and hyperparameter tuning are customized to the unique characteristics of each cryptocurrency.

The configuration and computational details for finetuning of our proposed ATCPHX model ATCPHX model consist of two steps: feature selection and hyperparameter tuning. These processes are formally defined in Algorithm 1 and Algorithm 2, respectively. The runtime analysis is discussed in the subsequent paragraph.This structured ensures the model's optimal performance. Below, we summarize the key methodological steps:

1. **Feature Selection:**
   (a) Spearman's correlation coefficient is used to identify the most relevant features for predicting the price trend of each cryptocurrency. This individualized approach captures the unique statistical properties and relationships inherent to each asset.
   (b) Features with a correlation coefficient below 0.4 are excluded from training. This selective process enhances predictive capability by retaining only the most impactful variables.
2. **Hyperparameter Tuning:**
   (a) The GridSearchCV method fine-tunes the hyperparameters of the XGBoost model. It systematically explores a predefined grid to `max_depth`, `learning_rate`, `n_estimators`, `subsample`, `min_child_weight`, and `gamma`.
   (b) Each cryptocurrency undergoes hyperparameter tuning individually, ensuring the model parameters are tailored to its specific data characteristics, thereby improving performance and prediction accuracy.

The time complexity of the hyperparameter optimization process is determined by three key factors: the number of cryptocurrencies n, the number of hyperparameter combinations k, and the time t required to train a single XGBoost model configuration. Due to the exhaustive nature of GridSearch, which evaluates all k hyperparameter combinations for each cryptocurrency, the total computational complexity is $O(n \times k \times t)$. While this approach is computationally intensive, it is feasible for our study given the manageable number of cryptocurrencies and the efficiency of the XGBoost algorithm.

The test data for each cryptocurrency is evaluated using its corresponding fine-tuned model to predict price trends over horizons of 1 to 10 days. In this study, the forecast horizon is treated as a hyperparameter, optimized individually for each cryptocurrency. For each asset, the model is evaluated across all possible horizons (1–10 days), and performance metrics–including Accuracy, Precision, Recall, F1-Score, and ROC-AUC–are computed. The optimal horizon d_c^* for cryptocurrency c is selected based on the highest aggregate performance, typically considering both accuracy and F1-Score. For example, Solana might achieve peak performance at 4 days, while MANA might perform best at 3 days. This adaptive approach ensures that the model is tailored to the unique temporal dynamics of each cryptocurrency. The complete procedure is formalized in Algorithm 3.

In the pseudocode outlined in Algorithm 3, Spearman's correlation is employed to eliminate irrelevant features, while SMOTE is utilized to mitigate class imbalance in the dataset. The XGBoost model is subsequently fine-tuned via GridSearchCV, which conducts an exhaustive search over a predefined hyperparameter space to identify the optimal configuration for model training. The computational complexity of this pipeline stems primarily from the iterative processes of feature selection, data balancing, and hyperparameter optimization. Each of these steps depends critically on the scale of the input, including the number of cryptocurrencies, the dimensionality of the feature set, and the size of the hyperparameter grid. Notably, the exhaustive nature of GridSearchCV substantially contributes to the overall computational burden, as it

Algorithm 3 ATCPHX Framework for Adaptive Cryptocurrency Trend Prediction

Require: Set of cryptocurrencies C, Feature set F, Target variable y
Ensure: Trained models M_c for each cryptocurrency $c \in C$, Optimal prediction horizon d_c^*

procedure ATCPHX(C, F, y)
  for each cryptocurrency $c \in C$ **do**
    $F_c \leftarrow \emptyset$ ▷ Initialize selected feature set for c
    for each feature $f \in F$ **do**
      $\rho \leftarrow$ SpearmanCorrelation(f, y_c) ▷ Compute correlation
      if $\rho \geq 0.4$ **then**
        $F_c \leftarrow F_c \cup \{f\}$ ▷ Retain feature if correlation is high
      end if
    end for
    $D_c \leftarrow$ SMOTE$(X_c[:, F_c], y_c)$ ▷ Apply SMOTE to balanced dataset
    $hyperparameter\_grid \leftarrow$ {predefined parameters}
    $model \leftarrow$ XGBoost()
    $grid\_search \leftarrow$ GridSearchCV$(model, hyperparameter_grid)$
    $grid\_search.fit(D_c, y_c)$
    $M_c \leftarrow grid\_search.best\_estimator\_$ ▷ Save the best model for c

    $best\_score \leftarrow -\infty$
    $d_c^* \leftarrow 0$
    for $d \in \{1, 2, \ldots, 10\}$ **do** ▷ Evaluate prediction horizon from 1 to 10 days
      $X_{\text{pred}} \leftarrow$ prepare_features(d)
      $\hat{y} \leftarrow M_c.predict(X_{\text{pred}})$
      $score \leftarrow$ accuracy$(\hat{y}, y_{\text{true}})$
      if $score > best\_score$ **then**
        $best\_score \leftarrow score$
        $d_c^* \leftarrow d$ ▷ Update optimal prediction day
      end if
    end for
    return M_c, d_c^*
  end for
end procedure

requires evaluating all possible hyperparameter combinations. Despite this complexity, the approach ensures a rigorous and optimized model training process tailored to the unique characteristics of each cryptocurrency.

4 Experimental Results

In this section, we first present the evaluation metrics used to assess the performance of the ATCPHX model. We then describe the datasets employed for cryptocurrency trend prediction, detailing the sources and nature of the data. Next, we outline the configuration and parameter tuning process for the XGBoost algorithm. Finally, we

Table 4: Evaluation Metrics for the ATCPHX Model

Metric	Description
$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$ $\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$ $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$ $\text{F1-Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$	Where: • TP = True Positives • TN = True Negatives • FP = False Positives • FN = False Negatives
ROC-AUC Score	Area under the Receiver Operating Characteristic curve: •$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}$ (True Positive Rate) •$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}$ (False Positive Rate) •$\text{AUC} = \int_0^1 \text{TPR}(t)dt$ where t is the threshold
$\text{MI}(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log \frac{p(x,y)}{p(x)p(y)}$	Mutual Information between variables X and Y: • $p(x,y)$: joint probability distribution • $p(x)$, $p(y)$: marginal probability distributions

report and discuss the experimental results, highlighting the model's effectiveness and comparative performance.

4.1 Evaluation Criteria of the ATCPHX Model

The evaluation of the ATCPHX model is conducted using two distinct categories of criteria. The first category comprises standard performance metrics that quantify the model's effectiveness in predicting cryptocurrency price trends. The second category employs SHAP (SHapley Additive exPlanations) values to interpret the model by determining the contribution and importance of individual features in the prediction process. This dual approach not only measures predictive accuracy but also provides insights into the relevance of the features selected via Spearman's correlation coefficient. The performance metrics used for quantitative evaluation are summarized in Table 4.

The metrics in Table 4 quantify the predictive accuracy and overall performance of the ATCPHX model and its hyperparameter selection process. In contrast, SHAP values provide a complementary perspective by evaluating the contribution and importance of each input feature to the model?s predictions. SHAP enhances interpretability by enabling a deeper understanding of how the model uses input features to make decisions. Specifically, SHAP analysis allows us to:

1. **Model Interpretability:** SHAP addresses a fundamental challenge in machine learning by providing clear insights into how complex models make decisions. It quantifies the contribution of each feature to individual predictions, making the model's behavior transparent and understandable.
2. **Feature Importance Analysis:** SHAP enables rigorous evaluation of feature importance, identifying which variables most significantly influence predictions. This is particularly valuable for feature selection, dimensionality reduction, and model refinement.

3. **Fairness and Bias Detection:** SHAP helps identify features that may introduce discrimination or bias into model predictions. By revealing whether sensitive attributes (e.g., demographic variables) unduly influence outcomes, it supports the development of fairer and more ethical models.

The mathematical formulation of SHAP values is inherently more complex than the standard evaluation metrics in Table 4. For a prediction model f and an input feature set X with N features, the SHAP value ϕ_i for feature i in a specific prediction is conceptually represented as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} [f(S \cup \{i\}) - f(S)] \tag{3}$$

where:

- S is a subset of features excluding i,
- $|S|$ is the size of subset S,
- $f(S)$ denotes the model prediction using only the features in S,
- The term $f(S \cup \{i\}) - f(S)$ represents the marginal contribution of feature i to the subset S,
- The weighting factor $\frac{|S|!(N-|S|-1)!}{N!}$ accounts for the number of possible permutations of the feature subsets.

The SHAP value ϕ_i for a feature i quantifies its specific contribution to a model's prediction compared to the average prediction over the dataset. This value is derived by considering all possible subsets $S \subseteq N \setminus \{i\}$ of features, and averaging the marginal contribution $f(S \cup \{i\}) - f(S)$ of feature i across all such subsets, weighted by the factor $\frac{|S|!(N-|S|-1)!}{N!}$. This weighting accounts for the number of permutations of each subset size $|S|$, ensuring a fair distribution of feature contributions. SHAP values provide consistent and locally accurate feature attributions: consistency guarantees that if a feature's contribution to the model output increases, its SHAP value does not decrease, and local accuracy ensures that the sum of all SHAP values for a given instance equals the difference between the model's prediction and the average model output. This formulation makes SHAP a powerful tool for interpreting complex models by offering a unified, theoretically grounded measure of feature importance.

4.2 Datasets

Our study selected Telegram channels as the primary data source for sentiment analysis due to their widespread use among cryptocurrency traders and enthusiasts. These channels provide a unique combination of in-depth discussions, expert insights, and real-time reactions to market events, offering a rich and timely representation of market sentiment. Unlike broader social media platforms, Telegram groups often focus specifically on cryptocurrency markets, including emerging tokens that may receive

Table 5: Telegram Comment Sentiment Distribution

#Comments	#Positive	#Negative	#Neutral
13,425	5,291	3,890	4,244

limited coverage elsewhere. This focused discourse ensures that our sentiment analysis captures a diverse and comprehensive view of market dynamics, thereby enhancing the ATCPHX model's capacity to predict trends across both established and emerging cryptocurrencies.

Data was collected from November 1, 2023, to March 10, 2024, encompassing Telegram comments, historical price data, and auxiliary economic indicators. This period was selected based on empirical studies indicating that shorter windows are more effective in highly volatile markets like cryptocurrencies, where social media-driven sentiment and external factors cause rapid price shifts [22, 23]. Extending the time frame does not substantially improve prediction accuracy and may introduce noise due to non-stationary market conditions. Table 5 summarizes the statistics of the collected Telegram comment data.

Telegram dataset, after collection and preprocessing, were injected into the HDRB algorithm for aspect extraction and sentiment analysis. As previously mentioned, the HDRB algorithm employs an LDA-Concept model for feature extraction. Feature polarity is then determined through transfer learning using a pre-trained RoBERTa network, augmented with a Gated Recurrent Unit (GRU) and an attention layer.

The second dataset comprises historical price data for trending cryptocurrencies. A list of daily trending assets was mined using the CoinGecko API, which identifies emerging cryptocurrencies based on market behavior and search intensity. From this list, 24 predominantly lesser-known and emerging cryptocurrencies were selected for investigation. Their historical Open-High-Low-Close-Volume (OHLCV) data was subsequently retrieved via the Yahoo Finance API. The complete list of cryptocurrencies analyzed in this study is presented in Table 6.

In this research, economic data, the third data category, shows the market's fear and greed for transactions with a particular cryptocurrency. This data is obtained by the CFGI.io site API or through the Pandas package `ta` tool. The data extracted from CFGI are listed in Table 7, along with explanations.

It is important to note that the economic indicators employed in this research are specifically tailored to cryptocurrency markets, as opposed to general macroeconomic variables such as inflation, interest rates, or global stock indices. All indicators were either sourced directly from the CFGI.io API or programmatically calculated using the `ta` (Technical Analysis) Python library. This suite of indicators includes measures like Price Score, Volatility, Volume, Impulse, aggregated technical analysis signals, Dominance, search-based Trends, whale wallet activity, and Order Book dynamics. Each feature is designed to directly capture the behavioral and transactional dynamics unique to digital asset markets, rendering them more sensitive to short-term fluctuations than traditional macroeconomic factors. This targeted focus ensures the model

Table 6: List of Cryptocurrencies and Sample Sizes in the Dataset

Ticker	Name	Samples	Ticker	Name	Samples
ADA	Cardano	1,218	**ETC**	Ethereum Classic	873
ALGO	Algorand	835	**ETH**	Ethereum	4,139
APE	ApeCoin	645	**FIL**	Filecoin	573
ARB	Arbitrum	502	**FTM**	Fantom	904
ATOM	Cosmos	1,403	**HNT**	Helium	703
AVAX	Avalanche	926	**LINK**	Chainlink	1,036
AXE	AXE	541	**LTC**	Litecoin	1,321
BCH	Bitcoin Cash	1,543	**MANA**	Decentraland	1,260
BNB	Binance Coin	1,807	**MATIC**	Polygon	1,439
BTC	Bitcoin	4,631	**MEM**	Memecoin	751
DOGE	Dogecoin	2,057	**SOL**	Solana	1,804
DOT	Polkadot	1,037	**SUI**	SUI Network	741

captures market-specific drivers of price trends and enhances the reproducibility of this study, as other researchers can utilize the same APIs and open-source tools.

Subsequently, feature selection was conducted by applying Spearman's rank correlation coefficient. Features exhibiting a correlation greater than 40% with the target variabl–the price change trend–were selected as independent variables for training the ATCPHX model. As previously stated, this feature engineering process was applied individually to each cryptocurrency. The most influential features across the dataset, along with their average correlation values, are presented in Table 8.

The total number of features without considering the text columns is twenty features. Obviously, without performing feature engineering, using all features can cause the model's performance to drop. We also utilized the Mutual Information (MI) metric to select features for Bitcoin and Ethereum, further supporting the results obtained with Spearman's correlation. Mutual Information (MI) measures the mutual dependence between two variables. It quantifies the amount of information obtained about one variable through another. Unlike Spearman's correlation, which only captures linear or monotonic relationships, MI can detect non-linear dependencies between variables. The higher the MI value, the more information one variable provides about another [24]. We applied the Mutual Information technique to our dataset and compared the results with those obtained from Spearman's correlation. The selected features from both methods were broadly consistent, indicating the robustness of our initial feature selection. However, MI did highlight a few additional features in some cryptocurrencies with non-linear relationships that were not captured by Spearman's correlation alone. These extra features were incorporated into the model, and we observed a slight improvement in the overall model performance. Table 9 presents the feature selection results for predicting Ethereum and Bitcoin using two techniques: MI and Pearson's correlation coefficient.

As evidenced by the results in Table 9, both *Data_Technical* and *Data_Social* exhibit higher Mutual Information (MI) scores than their corresponding Spearman's correlation coefficients. This indicates that these features share more complex, non-linear relationships with the target variable. The two methods also demonstrate

Table 7: Economic Features

Indicator	Description
Price Score	Measures the cryptocurrency's price trend, assigning a value from -1 (bearish) to +1 (bullish) to reflect prevailing market sentiment. This score gauges the general feeling of greed or fear based on the direction and duration of the price trend.
Price Volatility	Quantifies the degree of price fluctuation within a normalized range of 0 to 1 for a specified period. Higher volatility indicates a riskier market, amplifying sentiments of greed in bull markets and fear in bear markets.
Volume	Indicates the level of market activity by measuring trading volume, normalized from 0 to 1. An increase signifies heightened interest, strongly correlating with market sentiment?greed in rising markets and panic in declining ones.
Impulse	Assesses the strength and direction of price movement relative to historical values on a scale from -1 to +1. Strong bullish or bearish impulses suggest intensified feelings of greed or fear, respectively.
Technical Analysis	Incorporates 26 popular technical indicators from traditional markets, categorized into trend-following indicators and oscillators. Each indicator is averaged and weighted by popularity to composite a sentiment score.
Social Media	Employs AI-driven natural language processing to analyze sentiment from posts on platforms like Twitter and Reddit. The model distinguishes between organic sentiment and promotional content to determine overall market mood.
Dominance	Examines the asset's market capitalization share relative to the total crypto market. Shifts in dominance reflect changing investor preference between Bitcoin and altcoins, providing insight into sector-level sentiment and correlation strength.
Trends	Analyzes search volume data for cryptocurrency-related queries on major search engines, comparing current levels to historical baselines. Uses specific intent-based keywords (e.g., "buy X") rather than generic terms to gauge interest.
Whales	Tracks large transactions initiated by major investors ("whales"), specifically movements between personal wallets and exchanges. A higher ratio of crypto moving to exchanges suggests selling intent (greed), while movement away indicates accumulation (fear).
Order Books	Analyzes the limit order book depth across major exchanges to assess buying and selling pressure. A preponderance of buy orders near the price indicates bearish sentiment, while significant sell order depth suggests bullish sentiment.

divergence in their selection of other features; for instance, MI assigned a higher importance to the *Data_Orders* feature for Ethereum (ETH) than Spearman did. A key point of differentiation is the Volume feature, which was effectively rejected by the Spearman criterion but was selected by MI, albeit with a relatively low dependence

Table 8: Influential Features and Their Average Spearman Correlation Values

Feature	Open	High	Low	Close	Sentiment	Technical	Price	Impulse
`Spearman Value`	0.461	0.425	0.473	0.52	0.567	0.542	0.485	0.532

Table 9: Comparison of Feature Selection Metrics for BTC and ETH

Feature	Spearman (BTC)	MI (BTC)	Spearman (ETH)	MI (ETH)
Data_Technical	0.54	0.62	0.49	0.57
Data_Social	0.56	0.65	0.52	0.60
Data_Volume	0.47	0.51	0.45	0.50
Data_Impulse	0.50	0.58	0.53	0.60
Data_Dominance	0.42	0.47	0.39	0.46
Data_Trends	0.45	0.49	0.43	0.47
Data_Orders	0.49	0.52	0.48	0.53
Open	0.47	0.40	0.41	0.45
Close	0.48	0.42	0.46	0.43
Volume	0.38	0.43	0.36	0.41

percentage. The convergence of these results with the subsequent SHAP analysis supports the conclusion that the proposed hybrid feature selection method demonstrates robust performance.

Furthermore, to evaluate the predictive power of the selected features, a SHAP summary plot was employed for Bitcoin Cash (BCH), as illustrated in Figure 3. SHAP (SHapley Additive exPlanations) quantifies the marginal contribution of each feature to the model's predictions. In this visualization, the vertical axis lists the features ranked by their global importance, while the horizontal axis represents the SHAP value, which denotes the magnitude and direction (positive or negative) of each feature's impact on the model output for individual predictions. The SHAP summary plot in Figure 3 can be interpreted according to the following key points:

1. Each point on the chart represents the SHAP value for a single observation (data sample) from the dataset, illustrating the feature's contribution to the prediction for that specific instance.
2. The color gradient of the points (from red to blue) corresponds to the actual value of the feature for that observation. Red signifies a low feature value, while blue signifies a high feature value.
3. The horizontal position of a point indicates the direction and magnitude of its effect on the prediction. Points located to the right of the central zero line exert a positive effect (pushing the prediction higher), while points to the left exert a negative effect (pushing the prediction lower).
4. The overall importance of a feature is indicated by the dispersion of its points along the horizontal axis. Features with points spread far from zero (e.g., Attribute 4: Low) have a substantial impact on the model's output. For

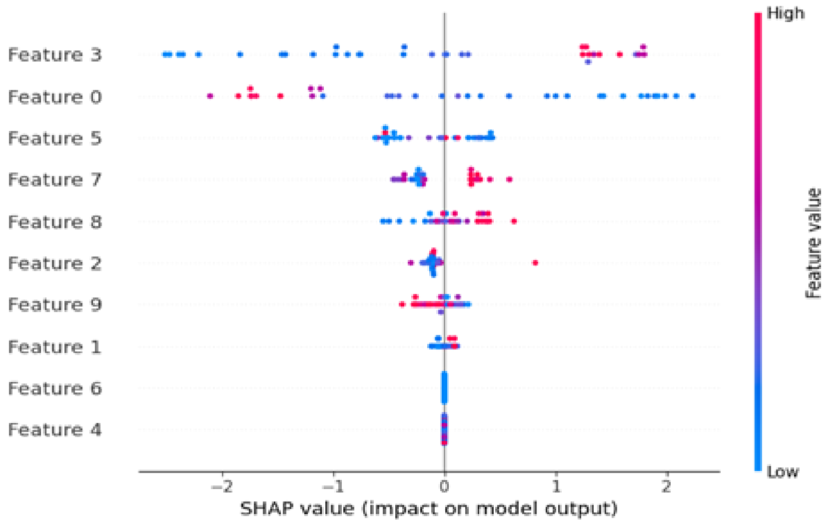

Fig. 3: SHAP chart to evaluate selected features to determine the trend of BCH

instance, Low exhibits the most significant negative impact, whereas Sentiment demonstrates the strongest positive effect.

5. The vertical line within the distribution of points for each feature represents the mean absolute SHAP value. A mean value closer to zero indicates a feature with lower overall impact on the model's predictions.

Based on the SHAP summary plot in Figure 3, where a positive SHAP value denotes a feature's positive contribution to the model's output, the features can be ranked from the most positively influential to the least as follows:

1. **Feature 3 (Sentiment)** exhibits the strongest positive influence on the model's output, as indicated by its top position on the vertical axis. The wide dispersion of its SHAP values signifies that its impact varies considerably across different predictions.
2. **Feature 0 (Impulse)** also demonstrates a substantial positive impact. The slightly lower spread of its values, compared to Feature 3, suggests a more consistent effect on the model's predictions.
3. **Feature 5 (Technical)** has a significant yet complex influence. The presence of both high positive and high negative SHAP values indicates that it can strongly affect predictions in either direction, depending on the context.
4. **Feature 7 (Close)** has a mixed and non-uniform impact on the prediction of Bitcoin Cash (BCH) trends. Its SHAP values are distributed across both the positive and negative domains, suggesting its effect is highly dependent on other market conditions.

5. **Feature 8 (Open)** demonstrates a moderate overall influence. Similar to the Close price, it contributes both positively and negatively to various predictions, indicating a contextual relationship with the target variable.

In light of the above,, a rigorous and systematic feature selection process was employed to ensure the inclusion of all relevant predictive factors for each individual cryptocurrency. This tripartite methodology was designed to capture a comprehensive range of statistical relationships:

1. **Pearson's Correlation Coefficient**: This measure identified features with strong linear relationships to the target variable. It served as the foundation for selecting statistically significant features that exhibit direct, linear correlations with price movements.
2. **Mutual Information (MI)**: To complement the linear approach, Mutual Information was employed to quantify non-linear dependencies between features and the target. This method is particularly adept at revealing complex, non-parametric interactions that traditional correlation measures might overlook.
3. **SHAP Values for Interpretability**: Beyond statistical selection, SHAP (SHapley Additive exPlanations) values were utilized to provide a model-agnostic interpretation of feature importance. This framework precisely quantifies the marginal contribution of each feature to the model's predictions, ensuring robustness and transparency by preventing any single feature from exerting disproportionate influence.

4.3 Hyperparameter Fine-Tuning of XGBoost

As previously stated, the XGBoost model was fine-tuned individually for each cryptocurrency. This approach is justified by the algorithm's high execution speed and the reusability of the stored model artifacts, making large-scale optimization feasible. The implementation was conducted in a Google Colab environment using Python 3, leveraging hardware resources including 13 GB of RAM and a T4 GPU.

The tuning process aimed to identify the optimal values for five key hyperparameters: `max_depth`, `learning_rate`, `n_estimators`, `subsample`, `min_child_weight`, and `gamma`. A Grid Search with 5-fold cross-validation (GridSearchCV) was employed for this purpose. Following methodologies established in prior literature [25, 26], three distinct values were evaluated for each hyperparameter (see Algorithm 2).

This configuration resulted in a total of $3^6 = 729$ unique candidate combinations. With 5 folds per candidate, the procedure required 3,645 complete model fits. Given an average fitting time of 0.1 seconds per model, the entire grid search process was completed in approximately 6.07 minutes per cryptocurrency. The resulting optimal hyperparameter configurations for Bitcoin, Atom, and MANA are presented in Table 10.

Table 10: Optimal XGBoost Hyperparameters for Selected Cryptocurrencies

Cryptocurrency	`max_depth`	`learning_rate`	`n_estimators`	`subsample`	`min_child_weight`	`gamma`
BTC	3	0.02	250	0.7	1	0.1
ATOM	3	0.02	300	0.7	3	0.1
MANA	4	0.02	300	0.9	1	0

Table 11: Accuracy Comparison of Predictive Models

Model	Cryptocurrency			
	ETH	ATOM	APE	AXE
XGBoost	0.653	0.672	0.597	0.581
LSTM	0.597	0.605	0.550	0.540
RF	0.631	0.650	0.551	0.539

4.4 Results of Cryptocurrency Trend Prediction

This section presents the performance evaluation results of the ATCPHX model over 300 execution epochs, assessed against the evaluation metrics outlined in Table 4. To validate the effectiveness of the XGBoost algorithm for cryptocurrency price trend prediction, a baseline comparison was conducted against Random Forest (RF) and Long Short-Term Memory (LSTM) models using accuracy (ACC) as the primary metric.

The LSTM architecture was configured with 64 LSTM units, followed by dense layers of 128 and 64 neurons, and a final output layer with a single neuron. The ReLU activation function was used for all intermediate layers, while the output layer utilized a sigmoid activator. The model employed L1-L2 kernel regularization, a learning rate of 0.001 (`learning_rate`=100), and incorporated EarlyStopping based on validation loss to mitigate overfitting. The Random Forest classifier was implemented with 100 estimators (`n_estimators`=0.001). For this comparative analysis, the XGBoost classifier utilized all default parameters as defined in the `xgboost` package.

The results of the model comparison for predicting the price trends of AX, APE, ATOM, and ETH cryptocurrencies, summarized in Table 11, demonstrate that the default XGBoost model–even prior to fine-tuning–outperforms both the tuned LSTM and Random Forest (RF) models. This finding confirms the superior performance of the XGBoost method for predicting cryptocurrency price trends, which is consistent with prior research [26–28] and is further supported by the comprehensive evaluation results in Table 12, which details the performance ofimplemented ATCPHX model across all 24 cryptocurrencies, evaluated against the ERM criteria.

Table 12 shows that the average accuracy of the model is equal to 0.703. The best accuracy in prediction related to LTC and BCH cryptocurrencies with 83% accuracy, the best recall related to AVAX cryptocurrency equal to 90%, the best precision rate related to BCH cryptocurrency equal to 89%, the best F1-Score value related to LTC

Table 12: Comprehensive Performance Eval. of the ATCPHX

Cryptocurrency	Evaluation Metric				
	Accuracy	Recall	Precision	F1-Score	ROC-AUC
ADA	0.770	0.882	0.714	0.789	0.770
ALGO	0.590	0.660	0.640	0.651	0.695
APE	0.770	0.916	0.610	0.730	0.836
ARB	0.770	0.882	0.714	0.789	**0.903**
ATOM	0.660	0.730	0.580	0.647	0.717
AVAX	0.740	**0.900**	0.750	0.818	0.784
AXE	0.694	0.875	0.608	0.717	0.768
BCH	**0.830**	0.809	**0.894**	0.850	0.860
BNB	0.594	0.842	0.581	0.688	0.749
BTC	0.660	0.894	0.629	0.739	0.755
DOGE	0.750	0.856	0.653	0.741	0.860
DOT	0.750	0.842	0.720	0.780	0.798
ETC	0.660	0.650	0.720	0.684	0.846
ETH	0.635	0.795	0.672	0.728	0.781
FIL	0.770	0.850	0.772	0.809	0.765
FTM	0.722	0.880	0.660	0.761	0.830
HNT	0.581	0.523	0.631	0.572	0.681
LINK	**0.830**	0.882	0.789	**0.830**	0.863
LTC	0.638	0.660	0.550	0.606	0.736
MANA	0.750	0.809	0.772	0.790	0.820
MATIC	0.610	0.789	0.600	0.681	0.712
MEM	0.823	0.882	0.789	**0.830**	0.896
SOL	0.694	0.550	0.846	0.660	0.831
SUI	0.600	0.560	0.823	0.660	0.740
Average	0.703	0.780	0.696	0.731	0.791

Table 13: Impact of SA on Predictive Performance

Cryptocurrency	Model Variant	Accuracy	F1-Score
BTC	Without Sentiment	0.610	0.580
	With Sentiment	0.660	0.739
ETH	Without Sentiment	0.590	0.550
	With Sentiment	0.635	0.728

and MEM cryptocurrencies equal to 83% and finally, the best ROC-AUC benchmark rate for ATOM cryptocurrency is equal to 90%.

Table 13 shows the performance results of the model with and without using the sentiment score for predicting BTC and ETH price trends.

Table 13 shows the performance results of the model with and without sentiment scores for predicting BTC and ETH price trends. The results demonstrate that incorporating sentiment scores improves both ACC and F1-Score for both cryptocurrencies. Specifically, for BTC, ACC increased from 0.61 to 0.66 and the F1-Score from 0.58 to 0.739. Similarly, for ETH, ACC rose from 0.59 to 0.635 and the F1-Score from 0.55

Table 14: Optimal Forecast Horizon (Day) to Predict Price Trends for Various Cryptocurrencies.

Forecast Horizon	Day									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Crypto	DOGE	FIL	ARB FTM LINK	ALGO DOT ETC HNT	ATOM BNB ETH MEM	ADA LTC MANA SOL	BCH BTC MATIC SUI	AXE	APE AVAX	–

to 0.728. These consistent improvements highlight the significant impact of sentiment analysis on enhancing the model's predictive performance.

The optimal prediction horizons identified for each cryptocurrency can indeed vary significantly across different market conditions or periods. Market conditions such as high volatility, bull or bear trends, and macroeconomic events can influence the optimal time frames for predicting price trends. The periods chosen for training and testing the model also play a crucial role. Historical data from different periods may exhibit varying patterns due to changes in market dynamics, regulatory developments, or technological advancements. For example, the behavior of cryptocurrencies in the early stages of their development might differ significantly from their behavior in more mature stages. As such, the model must be adaptable to these changes to maintain its accuracy and reliability. Also, the importance of choosing the period affecting the price trend of cryptocurrencies is consistent with the results of experiments by Rotman et al. [29]. As mentioned, the time frame of price trend prediction is considered a hyperparameter. Therefore, this interval is different depending on the best performance of the model for each cryptocurrency. Table 14 shows the best time interval for predicting the price trend of 24 cryptocurrencies.

As shown in Table 14, the optimal lag between shifts in market sentiment and their full effect on price changes is between three and seven days. In other words, the impact of disseminated news and the resulting market sentiment of fear and greed typically peaks within this 3-to-7-day window. By leveraging this known lag period to predict price trends, an investor can capitalize on potential gains or preserve capital by avoiding periods of heightened risk and potential losses. To contextualize the performance metrics presented in Table 12, Figure 4 illustrates the historical price changes for two established cryptocurrencies, SOL and BNB, alongside two emerging ones, SUI and HNT.

In Figure 4, the origin points are marked in green, while the optimal time intervals identified by the ATCPHX model are indicated by red points. The figure demonstrates the model's capability to accurately identify both rising and falling market peaks. For instance, an investor who purchases SOL or SUI on the final day indicated by the model could sell at the subsequent price peak–after 6 or 7 days, respectively–to secure a profit. Furthermore, Table 15 provides a comparative analysis of the ATCPHX model's performance against the following existing studies:

(a) SOL (Solana)

(b) BNB (Binance Coin)

(c) SUI (Sui)

(d) HNT (Helium)

Fig. 4: Historical price changes of four major cryptocurrencies. Data sourced from coinmarketcap.com.

1. Critien et al. [16] introduced a comprehensive model to predict Bitcoin price changes and trends by analyzing Twitter sentiment and tweet volumes. Their methodology leverages sentiment scores derived from tweets using the VADER tool and incorporates tweet volume as a feature to predict both the direction and magnitude of Bitcoin price movements. The study employs two main neural network models: a Recurrent Neural Network (RNN) to capture temporal dependencies and a Convolutional Neural Network (CNN) to identify spatial patterns in the data. Additionally, a novel approach is used to predict the magnitude of price changes by framing it as a multi-class classification problem. The models are evaluated on their ability to predict daily price changes, with a voting classifier combining the predictions for improved accuracy.
2. Valencia et al. [30] introduced an approach for predicting cryptocurrency market movements by leveraging machine learning techniques and sentiment analysis of Twitter data. They explored the predictive power of neural networks (NN),

support vector machines (SVM), and random forests (RF) using both market data and social data as inputs. Their methodology aimed to compare the effectiveness of these machine learning tools across four major cryptocurrencies: Bitcoin, Ethereum, Ripple, and Litecoin. By employing sentiment scores derived from Twitter alongside traditional market data, the study sought to understand the potential correlations between public sentiment and price movements within the highly volatile cryptocurrency market. The research highlighted the varying effectiveness of the different models and data types across the cryptocurrencies studied, ultimately demonstrating the feasibility of using machine learning and sentiment analysis for market prediction in the emerging digital currency landscape.

3. Garg et al. [31] introduced CrypTop12, a comprehensive dataset tailored for cryptocurrency price movement prediction, utilizing both Twitter sentiment and historical price data of the top 12 cryptocurrencies by market capitalization. This dataset aims to bridge the gap in existing research tools by providing a robust platform for analyzing the impact of social media sentiment and market trends on cryptocurrency prices. It encompasses over 576K tweets and historical price data spanning 1255 days, meticulously refined to emphasize tweets with significant relevance to price fluctuations. Additionally, the authors adapted baseline methods for price prediction, integrating modifications to suit the unique characteristics of the cryptocurrency market, such as continuous trading activity and the absence of adjusted closing prices. Through quantitative analysis and adaptation of existing models like Hybrid Attention Networks and StockNet, this study demonstrates the potential of CrypTop12 in advancing the predictive analytics of cryptocurrency price movements.
4. Zhou et al. [32] introduced a novel portfolio optimization model that integrates multi-source data, including historical cryptocurrency trading data, social media sentiment (from Twitter), and Google Trends, to predict cryptocurrency price movements. The model employs a Support Vector Machine (SVM) for price prediction, considering sentiment indicators derived from Twitter data and search interest from Google Trends as predictive variables. This approach significantly diverges from traditional portfolio strategies by incorporating both forecasted price movement information and the minimum variance model to optimize portfolio allocation.
5. Zou et al. [33] introduced a multimodal PreBit model to predict extreme Bitcoin price movements by integrating Twitter data with technical indicators and correlated asset prices. The model employs FinBERT embeddings to capture the financial context of tweets, combining them with candlestick data and various technical indicators in a Support Vector Machine (SVM) and Convolutional Neural Network (CNN) hybrid framework. The study demonstrates that incorporating detailed Twitter content significantly enhances prediction accuracy, utilizing a dataset of 5,000 daily tweets mentioning Bitcoin from 2015 to 2021. An ablation study of a trading strategy based on the model's predictions further validates its effectiveness in providing profitable and risk-mitigated trading signals.

6. Zhong et al. [34] introduced the LSTM-ReGAT model, a network-centric approach for predicting cryptocurrency price trends. The model combines long short-term memory (LSTM) for capturing individual cryptocurrency features with a relations graph attention network (ReGAT) to incorporate interrelations between cryptocurrencies. By constructing a cryptocurrency network based on technological foundations, industry affiliations, and investor co-attention, the model differentiates the importance of these relations through hierarchical attention mechanisms. The study's experimental analysis, using real-world market data, demonstrates the model's superior performance in predicting price trends and maximizing trading profits, offering significant insights for investment decision-making in the volatile cryptocurrency market.
7. Saleem et al. [35] introduced an innovative framework combining prospect theory and electronic word of mouth (eWOM) principles to forecast Bitcoin price fluctuations through Twitter sentiment analysis. The study collected over 3 million tweets from 2013 to 2022 using the keywords "bitcoin" and "btc," and incorporated normalized positive and negative sentiment scores and variance to enhance the sentiment analysis model. Their findings emphasize the significant impact of negative sentiments on driving Bitcoin price declines and underscore the importance of real-time monitoring of sentiment shifts.
8. Dag et al. [36] introduced a Tree Augmented Naïve Bayes (TAN) methodology designed to classify cryptocurrency trends, mainly focusing on Bitcoin (BTC) price movements. The methodology combines a data-driven approach with a four-step process: initial feature selection using Random Forest (RF), refinement through Simulated Annealing (SA), and the development of a TAN model that highlights conditional and interdependent relationships among variables influencing BTC prices. The model aims to aid short-term investors by providing a semi-interpretable and economical tool that balances accuracy with simplicity, achieving high-performance metrics such as an AUC of 0.652 and an accuracy of 0.667 with just six variables.
9. Ozer et al. [37] proposed a machine learning-based trading system that utilizes various classifiers, including Logistic Regression, SVM, Random Forest, and XGBoost, to predict the price movements of cryptocurrencies such as Bitcoin, Ethereum, and Litecoin. The study focuses on different time frames, specifically 4-hour and 1-day intervals, to enhance prediction accuracy. The novelty of their approach lies in integrating a robust outlier detection mechanism and optimizing the classifiers for short-term trading strategies. The results demonstrate that the proposed system, mainly when using Random Forest and Logistic Regression, outperforms traditional trading strategies, achieving higher returns on investment (ROI) and reducing maximum drawdown across the evaluated cryptocurrencies.
10. Dikovitsky [38] introduces a cascade classifier model that evaluates the strength of news impact (strong or weak) on Bitcoin price in two steps and, if considered strong, predicts the price movement (increase or decrease). This study uses three models: GloVe (as a simpler, lighter baseline), BERT (to better understand the text's meaning), and GPT with 1.5 billion parameters (a larger model with an increased ability to predict the direction of price movement).

Table 15: Performance comparison of the ATCPHX model against state-of-the-art methods. Highest values for each metric are highlighted in bold.

Model	**ACC**	**Recall**	**Precision**	**F1-Score**	**AUC-ROC**
Critien et al. (2022) [16]	BTC: 0.63	–	–	–	–
Valencia et al. (2019) [30]	BTC: **0.72** (±0.06) ETH: 0.61 XRP: 0.61 (±0.04) Average: 0.646	0.76 0.37 0.62 0.583	0.72 0.61 0.61 0.646	**0.739** 0.46 **0.60** 0.593	–
Garg et al. (2021) [31]	12 crypto (AVG): 0.56	-	-	0.51	–
Zhou et al. (2023) [32]	BTC: 0.61 ETH: 0.61 XRP: **0.72** ADA: 0.58 DOGE: 0.72 DOT: 0.67 LTC: 0.61 Average: 0.65	0.61 0.61 0.72 0.58 0.72 0.67 0.61 0.65	0.65 0.63 **0.81** 0.58 **0.81** **0.80** **0.64** 0.70	0.60 0.60 **0.66** 0.58 0.66 0.62 0.59 0.62	–
ATCPHX	BTC: 0.68 ETH: **0.635** ADA: **0.77** DOGE: **0.75** DOT: **0.75** LTC: **0.638** Average: 0.703	**0.894** **0.795** 0.882 0.852 0.842 0.66 0.772	0.629 **0.672** **0.714** 0.653 0.72 0.55 0.635	**0.739** **0.728** **0.789** **0.741** **0.78** **0.606** 0.731	**0.775** **0.781** **0.77** **0.86** **0.798** **0.736** 0.786
Zou et al. (2023) [33]	BTC: 0.61 (Avg in 4 Strategy)	0.59	0.62	0.604	–
Zhong et al. (2023) [34]	BTC: 0.629	0.604	0.499	0.546	0.661
Saleem et al. (2024) [35]	BTC: 0.55	0.55	0.59	0.56	–
Dag et al. (2023) [36]	BTC: 0.667	0.697	0.641	0.652	0.652
Ozer et al (2022)[37]	BTC: 0.565 ETH: 0.562 LTC: 0.0546 Average: 0.522	–	–	–	–
Dikovitsky (2025) [38]	BTC: 0.577	0.541	**0.775**	**0.73**	–

Table 15 demonstrates that the proposed ATCPHX model improves upon the accuracy of other models for most cryptocurrencies, with the notable exception of BTC trend prediction. More significantly, the ATCPHX model shows a substantial improvement in the F1-Score compared to all other studies. As the F1-Score is the harmonic mean of Precision and Recall, its enhancement indicates a balanced and simultaneous improvement in both of these metrics. This suggests that the ATCPHX model is robust and stable, effectively identifying both positive and negative trends without a specific bias. Furthermore, the comparative analysis of average accuracy and F1-score across models, illustrated in Figure 5, provides additional visual evidence of ATCPHX's superior performance.

The results in Figure 5 demonstrate that the ATCPHX model has achieved stable and high performance in both accuracy and F1-Score. This balanced performance

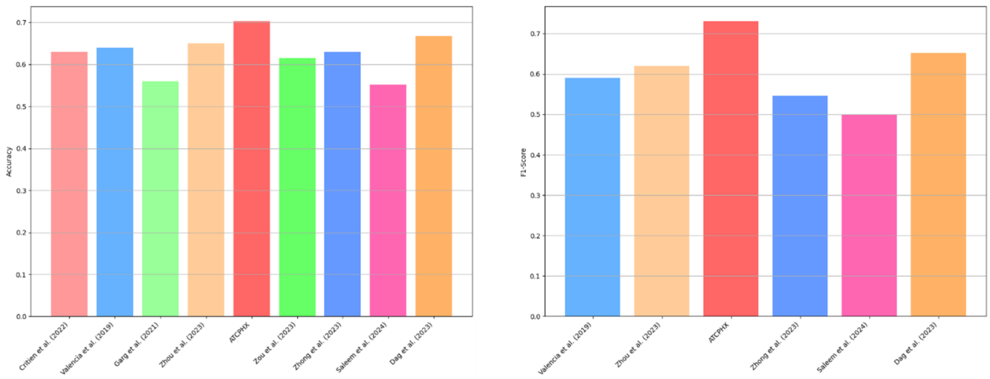

Fig. 5: Average Accuracy and F1-Score comparison for the ATCPHX model and state-of-the-art models.

Table 16: Comparison of cryptocurrency price trend prediction performance using confusion matrix analysis.

Crypto	Value	T	F
BTC	P	13 (TP)	3 (FP)
	N	2 (TN)	2 (FN)
ETH	P	14	2
	N	1	3
MANA	P	15	1
	N	3	1
DOT	P	14	1
	N	1	4

indicates that the model maintains good predictive accuracy while also achieving significant improvements in precision and recall. In contrast, studies by Valencia et al. [30] and Zhong et al. [34] achieved higher accuracy at the expense of a lower F1-Score, suggesting a decline in either recall or precision.

An important aspect to investigate is the model's error distribution in predicting upward versus downward trends. Therefore, Table 16 presents a detailed analysis of trend forecasting performance for BTC, ETH, MANA, and DOT since the beginning of December 2023. We employed a confusion matrix with Positive (P) and Negative (N) categories, using True (T) and False (F) values to evaluate prediction performance [33].

Table 16 shows the performance of the prediction model for four cryptocurrencies (BTC, ETH, MANA, and DOT) by indicating the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each cryptocurrency. For Bitcoin (BTC), the model demonstrates a relatively high number of true positives, indicating good performance in predicting upward trends. However, the instances of

Table 17: Return on investment (ROI) comparison between the proposed ATCPHX model and baseline machine learning models [37] for a 3-month trading period.

Cryptocurrency	Model	ROI (1D, 3-Month) %
BTC	**ATCPHX (AVG)**	**74**
	Logistic Regression	65
	SVM	73
	Random Forest	51
	CatBoost	57
ETH	ATCPHX (AVG)	95
	Logistic Regression	77
	SVM	**106**
	Random Forest	84
	CatBoost	83
LTC	**ATCPHX (AVG)**	**101**
	Logistic Regression	69
	SVM	94
	Random Forest	47
	CatBoost	67

Note: Best performance for each cryptocurrency is highlighted in bold.

false positives and false negatives suggest some misclassification in both upward and downward trend predictions. For Ethereum (ETH), the model also shows many true positives, reflecting strong predictive capability for upward trends. Nevertheless, the noticeable presence of false negatives indicates some difficulty in accurately predicting downward trends. This pattern suggests that while the model effectively identifies when ETH is likely to increase in value, it is less accurate in predicting decreases. In the case of MANA, the model performs exceptionally well, with a high count of true positives and minimal false positives and false negatives. This result indicates robust predictive performance for both upward and downward trends, showcasing the model's substantial accuracy for this cryptocurrency. For Polkadot (DOT), the model performs well in predicting upward trends, as evidenced by the high number of true positives. However, it struggles more with downward trends, as indicated by the higher count of false negatives. This discrepancy points to a potential need for further refinement in predicting when DOT is likely to decrease in value. Furthermore, Table 17 compares the effectiveness of the ATCPHX model in terms of ROI criteria with the approach by Ozer et al. [37].

In Table 17, the performance of the ATCPHX model is compared with other commonly used machine learning models, including Logistic Regression, SVM, Random Forest, and CatBoost, based on ROI over a three-month period. The values presented for the ATCPHX model represent averages calculated from the best-performing periods within these three months.

For BTC, the ATCPHX model achieved a higher average ROI of 74%, outperforming all other models, with SVM being the closest competitor at 73%. This suggests that ATCPHX has strong predictive capabilities and provides more consistent profitability. Logistic Regression, with an ROI of 65%, underperformed compared to both ATCPHX and SVM, indicating that ATCPHX is better suited to capturing short-term fluctuations in Bitcoin's price. Random Forest and CatBoost showed significantly lower ROIs of 51% and 57%, respectively, highlighting that these ensemble methods may not be as effective as ATCPHX and SVM in this specific scenario.

For ETH, the ATCPHX model also performed well, achieving an ROI of 95%, although SVM achieved a slightly higher ROI of 106%. The performance gap between ATCPHX and Logistic Regression, which recorded an ROI of 77%, further indicates ATCPHX's superiority in adapting to market dynamics. While Random Forest and CatBoost performed relatively well, with ROIs of 84% and 83% respectively, they still lagged behind SVM, reaffirming the need for sophisticated models like ATCPHX to achieve optimal returns.

For LTC, the ATCPHX model demonstrated exceptional performance with an ROI of 101%, outperforming all other models. This result underscores the model's robustness across different cryptocurrencies. The substantial gap between ATCPHX and Logistic Regression, which achieved only a 69% ROI, suggests that Logistic Regression may not fully capture the market trends influencing Litecoin's price. Although SVM performed well with a 94% ROI, it still fell short compared to ATCPHX. Random Forest and CatBoost, with ROIs of 47% and 67% respectively, performed the weakest, particularly Random Forest, which struggled with Litecoin's market characteristics during the observed period.

Overall, the ATCPHX model consistently demonstrated competitive or superior average ROI across BTC, ETH, and LTC, highlighting its effectiveness in short-term price prediction and market trend analysis. The significant difference in ROI between ATCPHX and models like Logistic Regression and Random Forest suggests that ATCPHX's underlying algorithmic enhancements and feature selection processes are better suited to capturing the nuances of market behavior in volatile environments. To better illustrate these comparative performances, Figure 6 shows the ROI of each model over the three-month period for Bitcoin, Ethereum, and Litecoin.

The experimental results demonstrate the significant potential of the ATCPHX model for cryptocurrency trend prediction. To contextualize these findings within practical applications, the key implications are summarized below, encompassing both the tangible benefits for market participants and the potential challenges requiring consideration.

1. **Tangible Benefits for Investors and Traders:**
   - 1.1. **Enhanced Decision-Making:** The model's improved predictive accuracy, averaging 70% with a robust F1-Score of 73%, provides investors with more reliable signals for strategic entry and exit points. This capability directly supports maximizing returns while mitigating potential losses.
   - 1.2. **Proactive Risk Management:** By accurately forecasting both upward and downward trends, the model enables traders to implement preemptive risk

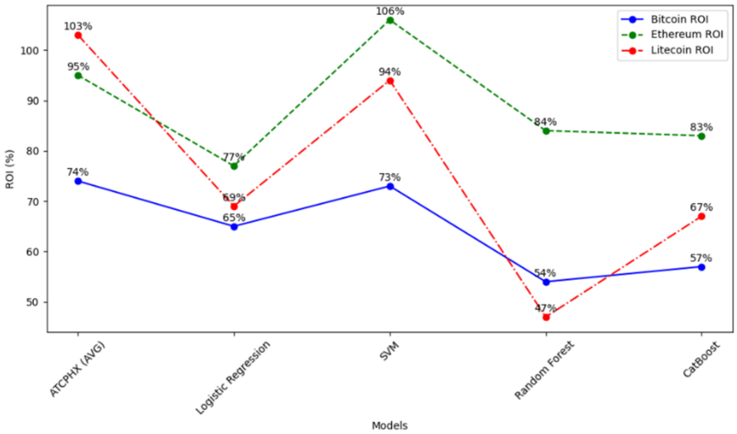

Fig. 6: Comparison of return on investment (ROI) between the ATCPHX model and the machine learning models proposed by Ozer et al. [37] for cryptocurrency trading.

management strategies. This includes exiting positions prior to anticipated downturns or capitalizing on emerging uptrends, thereby protecting capital and optimizing gains.

1.3. **Data-Driven Strategy Development:** The model's consistent performance allows for the development and rigorous back-testing of tailored trading strategies. Its customization for individual cryptocurrencies ensures that strategies are specifically aligned with the unique volatility and behavior of each asset.

1.4. **Increased Investor Confidence:** The demonstrated high accuracy and robustness of the ATCPHX model can strengthen confidence in algorithmic trading systems. This is particularly valuable for institutional investors when allocating substantial capital based on predictive analytics.

2. **Practical Challenges and Considerations:**

2.1. **Inherent Market Volatility:** Despite the model's design to account for volatility, extreme and unforeseen market events–such as regulatory announcements or macroeconomic shocks–can lead to deviations from predicted trends, representing a inherent risk.

2.2. **Dependence on Data Quality:** The model's performance is contingent on the availability, accuracy, and timeliness of input data, including social media sentiment and market indicators. Inconsistencies or delays in data feeds could directly impair prediction reliability.

2.3. **Algorithmic Implementation Risks:** An over-reliance on automated trading systems may lead to the neglect of fundamental analysis. Furthermore, technical execution risks, such as latency or system failures in high-frequency environments, pose additional operational challenges.

2.4. **Need for Continuous Adaptation:** Financial markets are dynamic and subject to evolving conditions. Maintaining prediction accuracy requires periodic model retraining and updates to incorporate new data patterns and market behaviors.

5 Conclusion

This study presented the ATCPHX model, a novel framework for predicting price trends of prominent cryptocurrencies by integrating multi-modal data sources: sentiment from Telegram social networks, historical market prices, and key economic indicators. Social media content was processed using the HDRB deep learning architecture to perform aspect-based sentiment analysis, capturing nuanced market sentiment. Feature selection was conducted for each cryptocurrency using Spearman's correlation analysis, followed by dataset balancing via the SMOTE technique to address class imbalance. The core of the ATCPHX framework utilizes a finely-tuned XGBoost algorithm, with hyperparameters optimized through GridSearch methodology specifically for each digital asset.

Comprehensive evaluation across 24 cryptocurrencies demonstrated the model's robust performance, achieving an average accuracy of 70% and an F1-Score of 73%. Furthermore, the model successfully identified optimal prediction horizons tailored to each cryptocurrency, providing valuable insights for strategic investment timing. The model's superior performance, particularly in comparison to established benchmarks, underscores the significant advantage of its personalized, multi-faceted approach to cryptocurrency trend forecasting.

Despite these promising results, this study acknowledges limitations including its primary focus on Telegram data and a constrained evaluation timeframe. Future research will expand this work by investigating the comparative impact of multiple social platforms?including Twitter, Reddit, and Facebook?across different cryptocurrencies, ultimately selecting optimal data sources on a per-asset basis to further enhance the personalized modeling approach introduced in this study.

Availability of data and materials

The datasets supporting the findings of this study are available from the corresponding author upon reasonable request.

Abbreviations

ABSA	Aspect-Based Sentiment Analysis
ACC	Accuracy
ATCPHX	Adaptive Trend Crypto Prediction with HDRB and XGBoost
BiGRU	Bidirectional Gated Recurrent Unit
CNN	Convolutional Neural Network
Concept-LDA	Concept-Latent Dirichlet Allocation
DLCFS	Deep Learning Cryptocurrency Forecasting considering Sentiment
FN	False Negative

FP	False Positive
HDRB	Hybrid Deep neural network model of RoBERTa and Bidirectional GRU
LSTM	Long Short-Term Memory
MI	Mutual Information
NN	Neural Network
OHLCV	Open, High, Low, Close, and Volume
P	Positive
N	Negative
RF	Random Forest
RNN	Recurrent Neural Network
ROI	Return On Investment
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
T	True
TN	True Negative
TP	True Positive
	Cryptocurrency Tickers:
ADA	Cardano
ALGO	Algorand
APE	ApeCoin
ARB	Arbitrum
ATOM	Cosmos
AVAX	Avalanche
BCH	Bitcoin Cash
BNB	Binance Coin
BTC	Bitcoin
DOGE	Dogecoin
DOT	Polkadot
ETC	Ethereum Classic
ETH	Ethereum
FIL	Filecoin
FTM	Fantom
HNT	Helium
LINK	Chainlink
LTC	Litecoin
MANA	Decentraland
MATIC	Polygon
MEM	Memecoin
SOL	Solana
SUI	SUI Network

References

[1] Khedr, A.M., Arif, I., El-Bannany, M., Alhashmi, S.M., Sreedharan, M.: Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management **28**(1), 3–34 (2021) https://doi.org/10.1002/isaf.1485

[2] Guo, T., Bifet, A., Antulov-Fantulin, N.: Bitcoin volatility forecasting with a glimpse into buy and sell orders. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 989–994 (2018). https://doi.org/10.1109/ICDM.2018.00124 . IEEE

[3] Jahanbin, K., Chahooki, M.A.Z.: Aspect-based sentiment analysis of twitter influencers to predict the trend of cryptocurrencies based on hybrid deep transfer learning models. IEEE Access **11**, 121656–121670 (2023) https://doi.org/10.1109/ACCESS.2023.3327660

[4] Mangalathu, S., Hwang, S.-H., Jeon, J.-S.: Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach. Engineering Structures **219**, 110927 (2020) https://doi.org/10.1016/j.engstruct.2020.110927

[5] Poongodi, M., Nguyen, T.N., Hamdi, M., Cengiz, K.: Global cryptocurrency trend prediction using social media. Information Processing & Management **58**(6), 102708 (2021) https://doi.org/10.1016/j.ipm.2021.102708

[6] Chen, Z., Li, C., Sun, W.: Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics **365**, 112395 (2020) https://doi.org/10.1016/j.cam.2019.112395

[7] Poongodi, M., Vijayakumar, V., Chilamkurti, N.: Bitcoin price prediction using arima model. International Journal of Internet Technology and Secured Transactions **10**(4), 396–406 (2020) https://doi.org/10.1504/IJITST.2020.108948

[8] Patel, M.M., Tanwar, S., Gupta, R., Kumar, N.: A deep learning-based cryptocurrency price prediction scheme for financial institutions. Journal of Information Security and Applications **55**, 102583 (2020) https://doi.org/10.1016/j.jisa.2020.102583

[9] Vo, A.-D., Nguyen, Q.-P., Ock, C.-Y.: Sentiment analysis of news for effective cryptocurrency price prediction. International Journal of Knowledge Engineering **5**(2), 47–52 (2019)

[10] Davchev, J., Mishev, K., Vodenska, I., Chitkushev, L., Trajanov, D.: Bitcoin price prediction using transfer learning on financial micro-blogs. In: The 16th Annual International Conference on Computer Science and Education in Computer

Science (2020)

[11] Abraham, J., Higdon, D., Nelson, J., Ibarra, J.: Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review **1**(3), 1 (2018)

[12] Li, L., Arab, A., Liu, J., Liu, J., Han, Z.: Bitcoin options pricing using lstm-based prediction model and blockchain statistics. In: 2019 IEEE International Conference on Blockchain (Blockchain), pp. 67–74 (2019). https://doi.org/10.1109/Blockchain.2019.00017 . IEEE

[13] Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z., Zhang, J.: Lstm based sentiment analysis for cryptocurrency prediction. In: Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, China, April 11–14, 2021, Proceedings, Part III 26, pp. 617–621 (2021). https://doi.org/10.1007/978-3-030-73200-4_42 . Springer

[14] Wołk, K.: Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Systems **37**(2), 12493 (2020) https://doi.org/10.1111/exsy.12493

[15] Amirshahi, B., Lahmiri, S.: Investigating the effectiveness of twitter sentiment in cryptocurrency close price prediction by using deep learning. Expert Systems, 13428 (2023) https://doi.org/10.1111/exsy.13428

[16] Critien, J.V., Gatt, A., Ellul, J.: Bitcoin price change and trend prediction through twitter sentiment and data volume. Financial Innovation **8**(1), 1–20 (2022) https://doi.org/10.1186/s40854-021-00313-6

[17] Parekh, R., Patel, N.P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I.E., Sharma, R.: Dl-guess: Deep learning and sentiment analysis-based cryptocurrency price prediction. IEEE Access **10**, 35398–35409 (2022) https://doi.org/10.1109/ACCESS.2022.3163501

[18] Low, J.M., Tan, Z.J., Tang, T.Y., Salleh, N.M.: Deep learning and sentiment analysis-based cryptocurrency price prediction. In: International Visual Informatics Conference, pp. 40–51 (2023). https://doi.org/10.1007/978-3-031-43148-7_4 . Springer

[19] Roy, P.K., Kumar, A., Singh, A., Sangaiah, A.K.: Forecasting bitcoin prices using deep learning for consumer centric industrial applications. IEEE Transactions on Consumer Electronics (2023) https://doi.org/10.1109/TCE.2023.3266318

[20] Chauhan, U., Shah, A.: Topic modeling using latent dirichlet allocation: A survey. ACM Computing Surveys (CSUR) **54**(7), 1–35 (2021) https://doi.org/10.1145/3462478

[21] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[22] Poudel, S., Paudyal, R., Cankaya, B., Sterlingsdottir, N., Murphy, M., Pandey, S., Vargas, J., Poudel, K.: Cryptocurrency price and volatility predictions with machine learning. Journal of Marketing Analytics **11**(4), 642–660 (2023) https://doi.org/10.1057/s41270-023-00238-2

[23] Ammer, M.A., Aldhyani, T.H.: Deep learning algorithm to predict cryptocurrency fluctuation prices: Increasing investment awareness. Electronics **11**(15), 2349 (2022) https://doi.org/10.3390/electronics11152349

[24] Zhou, H., Wang, X., Zhu, R.: Feature selection based on mutual information with correlation coefficient. Applied Intelligence **52**(5), 5457–5474 (2022) https://doi.org/10.1007/s10489-021-02635-5

[25] Drahokoupil, J.: Application of the XGBoost algorithm and Bayesian optimization for the Bitcoin price prediction during the COVID-19 period. FFA Working Papers 4.006, Prague University of Economics and Business (March 2022). Revised 09 May 2022. https://ideas.repec.org/p/prg/jnlwps/v4y2022id4.006.html

[26] Zhang, L., Zhou, R., Liu, Q., Xu, J., Liu, C., Babar, M.A.: Enhancing bitcoin transaction confirmation prediction: a hybrid model combining neural networks and xgboost. World Wide Web, 1–19 (2023) https://doi.org/10.1007/s11280-023-01178-8

[27] Shahbazi, Z., Byun, Y.-C.: Knowledge discovery on cryptocurrency exchange rate prediction using machine learning pipelines. Sensors **22**(5), 1740 (2022) https://doi.org/10.3390/s22051740

[28] Tiwari, R.G., Agarwal, A.K., Kaushal, R.K., Kumar, N.: Prophetic analysis of bitcoin price using machine learning approaches. In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 428–432 (2021). https://doi.org/10.1109/ISPCC53510.2021.9609394 . IEEE

[29] Rothman, T., Yakar, C.J.: Empirical analysis towards the effect of social media on cryptocurrency price and volume. European Scientific Journal **15**(31), 52–68 (2019) https://doi.org/10.19044/esj.2019.v15n31p52

[30] Valencia, F., Gómez-Espinosa, A., Valdés-Aguirre, B.: Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy **21**(6), 589 (2019) https://doi.org/10.3390/e21060589

[31] Garg, A., Shah, T., Jain, V.K., Sharma, R.: Cryptop12: A dataset for cryptocurrency price movement prediction from tweets and historical prices. In: 2021 20th

IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 379–384 (2021). https://doi.org/10.1109/ICMLA52953.2021.00064 . IEEE

[32] Zhou, Z., Song, Z., Xiao, H., Ren, T.: Multi-source data driven cryptocurrency price movement prediction and portfolio optimization. Expert Systems with Applications **219**, 119600 (2023) https://doi.org/10.1016/j.eswa.2023.119600

[33] Zou, Y., Herremans, D.: Prebit—a multimodal model with twitter finbert embeddings for extreme price movement prediction of bitcoin. Expert Systems with Applications **233**, 120838 (2023) https://doi.org/10.1016/j.eswa.2023.120838

[34] Zhong, C., Du, W., Xu, W., Huang, Q., Zhao, Y., Wang, M.: Lstm-regat: A network-centric approach for cryptocurrency price trend prediction. Decision Support Systems **169**, 113955 (2023) https://doi.org/10.1016/j.dss.2023.113955

[35] Saleem, T., Yaqub, U., Zaman, S.: Twitter sentiment analysis and bitcoin price forecasting: implications for financial risk management. The Journal of Risk Finance **25**(3), 407–421 (2024) https://doi.org/10.1108/JRF-07-2023-0166

[36] Dag, A., Dag, A.Z., Asilkalkan, A., Simsek, S., Delen, D.: A tree augmented naïve bayes-based methodology for classifying cryptocurrency trends. Journal of Business Research **156**, 113522 (2023) https://doi.org/10.1016/j.jbusres.2022.113522

[37] Ozer, F., Sakar, C.O.: An automated cryptocurrency trading system based on the detection of unusual price movements with a time-series clustering-based approach. Expert Systems with Applications **200**, 117017 (2022) https://doi.org/10.1016/j.eswa.2022.117017

[38] Dikovitsky, V.: Short-term cryptocurrency price forecasting based on news headline analysis. Frontiers in Blockchain **8**, 1627769 (2025) https://doi.org/10.3389/fbloc.2025.1627769