Forecasting the Prices of diverse cryptocurrencies Using Machine Learning, Deep Learning, and Explainable AI with Hybrid Sentiment Analysis from News Articles and Financial Market Data

MSc Research Project
Artificial Intelligence

Salman Ahmed
Student ID: 23379987

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .... Salman Ahmed
....

Student ID: ....23379987...

Programme: .... Artificial Intelligence .... **Year:** ......2024-25.....

Module: .... MSc Research Project ....

Supervisor: ...... Anderson Simiscuka ....

Submission Due Date: ....11/08/2024..

Project Title: ... Forecasting the Prices of diverse cryptocurrencies Using Machine Learning, Deep Learning, and Explainable AI with Hybrid Sentiment Analysis from News Articles and Financial Market Data ...

....

Word Count: ~8000 **Page Count**......21............

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ....Salman Ahmed..

Date: ....10/08/2024..

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	

Date:	
Penalty Applied (if applicable):	

Forecasting the Prices of diverse cryptocurrencies Using Machine Learning, Deep Learning, and Explainable AI with Hybrid Sentiment Analysis from News Articles and Financial Market Data

Salman Ahmed
23379987

Abstract

The cryptocurrency market is one of the most volatile financial spaces, mainly attributed to its exposure to speculative sentiments, lack of regulatory oversight, and unpredictable pricing factors. This study addresses the challenge of forecasting price movements by analysing ten major cryptocurrencies ETH, BTC, ADA, SOL, XRP, AVAX, LINK, DOT, LTC, and TRX using a combination of financial indicators and hybrid sentiment features derived from multiple online sources. The market data was collected via the Kraken API, while news and discussion-based sentiment signals were collected from Reddit, GDELT, and Mediastack using VADER, TextBlob, and FinBERT models. The derived sentiment scores were normalized, aggregated, and combined with technical indicators to formulate an integrated feature set. This feature set was then used to train six different machine learning models: Random Forest, Extreme Gradient Boosting, CatBoost, LightGBM, HistGradientBoosting, and Multi-Layer Perceptron Regressor. The model performances were evaluated using RMSE, MAE, and R^2 for each asset. Among the models analysed, tree- based ensemble methods namely LightGBM and CatBoost showed consistently better predictive power, with various models registering R^2 values greater than 0.98. To make the models more interpretable and support decision-making processes, Explainable Artificial Intelligence (XAI) techniques were applied. SHAP (Shapley Additive Explanations), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) were used to visualize and quantify the contribution of hybrid sentiment, volatility, and volume-based features for different assets. The results confirm that hybrid sentiment significantly enhances short-term forecasting accuracy, particularly for altcoins with considerable social influence. The proposed framework demonstrates that the combination of multi-source sentiment analysis and explainable machine learning allows for more transparent, interpretable, and accurate forecasting in complex digital asset markets.

1 Introduction

Cryptocurrencies, especially Bitcoin and Ethereum, have transformed finance markets owing to their decentralized frameworks, economic autonomy, and enhancement to an autonomous international financial community. Despite their growing frequency, their erratic pricing has caused market volatility, eroding investor confidence. Recent advances in ML, DL, natural language processing (NLP), and image processing have found wider uses in many industries [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. This paper examines multi cryptocurrencies data in combination with investor activity to train machine learning (ML) and deep learning (DL) models, aiming to identify the key factors influencing its price and volatility. The end aim is to create a stable and reliable market that inspires confidence among investors, financial analysts, and institutions. Market analysis indicates that pricing is dominated by investor quantity, with worldwide news and discussions about cryptocurrencies playing a significant role in investor decisions. As such, this piece of work integrates elements like social media updates, international news, together with financial indicators (e.g., opening price, closing price, low price, trading volume) to forecast cryptos price. The study also fills a gap in existing literature as it incorporates hybrid sentiment analysis combined with financial data to forecast cryptos price. It examines the efficiency of models like Random Forest, Extreme Gradient Boosting, Cat Boost, Light GBM, and Hist Gradient Boosting for predicting cryptos price using sentiment scores of international news, Reddit forums, and stock data. Additionally, it increases model transparency by utilizing explainable AI (SHAP) to interpret interactions between features in the best performing model. The paper highlights the role played by social media sentiment and financial metrics in prediction. The price of crypto currencies and market fluctuations, illustrating the worth of combining Using sentiment analysis with machine learning algorithms to improve price predictions. The organization of this manuscript is outlined as follows: Section 2 clarifies the related work analysis. Section 3 outlines the general methodology followed in This research paper outlines the algorithmic structure and system design in Section 4. Section 5 discusses the implementation of both deep learning and machine learning models, along with the metrics that were used for assessment. Section 6 analyses the experimental results and provides a critical analysis. Finally, Section 7 presents summarize the conclusion and delineates possible avenues for future research.

2 Related Work

In this section we will discuss critical analysis of recent and related literature, which is relevant to our work, evaluate them, dataset, algorithms, methodologies, evaluation metrics, and machine learning and deep learning models, and methodologies. The given analysis is categorised into subsections based on the phases of our implementation.

2.1 Sentimental Analysis

This section analyses sentiments in data from Twitter, Reddit, and news forums, associated with cryptocurrencies prices. Multiple studies have been done on sentiment analysis of text data for price prediction. In [16], a combination of GRU and Bi-LSTM is implemented to process data from CoinDesk, Twitter, and Coin Market Cap. By combining the sentiment scores of VADER and BERT, the model produces an average MSE of
0.024 for Bitcoin and 0.064 for Ethereum.
In [17], a hybrid model for sentiment classification that combines Decision Tree and Support Vector Machine. With an accuracy of 0.7975, the study examines financial news headlines and the stock price information that goes along with them.
In [22], cryptocurrency volatility was predicted using Twitter data and VADER sentiment analysis. The accuracy of the resulting Vector Autoregression model was 99.67%. [18] used TextBlob sentiment analysis to investigate sentiment-based stock price forecasts, emphasizing the slight influence of sentiment scores on price projections. The DistilBERT model demonstrated 98% accuracy in price predictions when BERT was employed for sentiment analysis on news headlines in the study in [19], [20]. In a similar vein, [21] used the CryptoBERT model with a 92% overall accuracy rate to forecast cryptocurrency values.

2.2 Explainable AI and Machine Learning for Cryptocurrency

Machine learning models' accountability and transparency can be improved via Explainable AI (XAI). To determine the factors that influence Bitcoin's price, especially during the Russia-Ukraine war, SHAP was used to Random Forest and Gradient Boosting models in [22]. In a similar vein, [23] examined feature importance for cryptocurrency price prediction using SHAP and contrasted supervised and unsupervised machine learning models.
While [24] used SHAP to increase investor trust by elucidating the mechanisms driving bitcoin prices across 21 distinct cryptocurrencies, [25] concentrated on showing XAI results in sentiment analysis models.

2.3 Embedding/Text Representation

Text embedding techniques like GloVe, BERT, and TF-IDF are tested for forecasting the prices of cryptocurrencies and financial markets. [26] shown more accuracy than alternative techniques while analyzing news headlines for gold price prediction using BERT embeddings. [27] outperformed conventional models in financial market forecasting by using FinBERT and BERT BoEC embeddings. The amount and quality of the dataset affect how well text embeddings work. On a smaller dataset of Bitcoin news headlines, TF-IDF outperformed BERT and FinBERT in [28], [29]. Research such as [3], [30] and [31] used Google Trends and Twitter data to forecast bitcoin prices, demonstrating that TF-IDF and BERT increased prediction accuracy. Table 1 provides an overview of the machine learning and deep learning techniques used to predict bitcoin prices.

2.4 Summary

This section emphasized obstacles in sentiment analysis, including subtle data features, domain-specific macro-indicators, and biases in sentiment analysis of unlabelled data. The employment of XAI approaches, text representations, and hybrid models is highlighted as vital for boosting prediction accuracy. The study intends to address these issues by utilizing cutting-edge techniques to optimize text representation, strengthen sentiment analysis, and increase model transparency all of which will ultimately lead to more precise predictions of bitcoin prices.

Table 1. *Summary of models implanted in related work*

Ref	Data Source	Models	Goal (Prediction)	Results
[32]	Twelvedata.com	CNN: LSTM, GRU, BiLSTM,	Crypto Price	LSTM - RMSE = 235.97,
[33]	Yahoo Finance	LSTM, kNN, LR, RR, DNN, RNN, RF	BNB, ETH, BTC price	Ridge - MSE = 0.000202
[34]	Twitter & ICICI	RF, LR, SVM, kNN	Stock Price	LR - 0.8741
[35]	Coin Market Cap	ARIMA, Trees, FbProphert.	ETH Price	ARIMA-produced less errors
[36]	Crypto Website Data	SVR, Integrated Learning	Crypto Price	0.783015776
[7]	NASDAQ Stock Exchange	LightGBM, XGBoost, Decision Trees, FL-LightGBM, CH L-LightGBM	Stock market	CHL-LightGBM achieved highest returns
[8]	Bitcoin via Api	XGBoost, GBR, RF, RF, SVR, NNR, Decision Tree, GPR , kNN, HGR	Bitcoin Price	GBR: $R^2 \approx 0.995$, RMSE = 0.01281, MAE = 0.005755
[10]	G-Research Crypto Forecasting	CatBoost, compared to Gradient Boosting, SVM, LR	Cryptocurrency market forecasting	CatBoost outperformed others: RMSP E improvement +12% over Gradient Boosting, +16.1%

				over SVM, +43.3% over LR
[11]	Cryptocurrency historical price data	Gradient Boost, KNN, SVR, LR, RF, AdaBoost, etc., using momentum indicators (RSI, MACD, ADX)	Cryptocurrency price	Gradient Boost achieved highest accuracy: $R^2 = 0.968$
[12]	Historical time series for Bitcoin & Ripple	XGBoost, AdaBoost, CatBoost	Long-term cryptocurrency value	AdaBoost & XGBoost
[13]	Kaggle dataset (2017–2022) for Bitcoin, Binance Coin, Ethereum, Tether (USDT)	KNN, Decision Tree, Random Forest, XGBoost, CatBoost	Closing price for 4 cryptocurrencies	RF & CatBoost best; RMSE — BTC: 365, BNB: 2.99, ETH: 29.99, USDT: 0.0023

3 Methodology

Using a combination of Reddit discussions, global news headlines, and Ethereum financial data, this study aims to forecast changes in the price of cryptos. For predictive modelling, it examines the relationship between sentiments and financial variables using a hybrid sentiment and financial research pipeline. To provide explanation and assessment, the study additionally highlights model transparency by investigating feature interactions and contributions with SHAP. The workflow used in the research study and the fundamental research technique are depicted in Fig. 1.

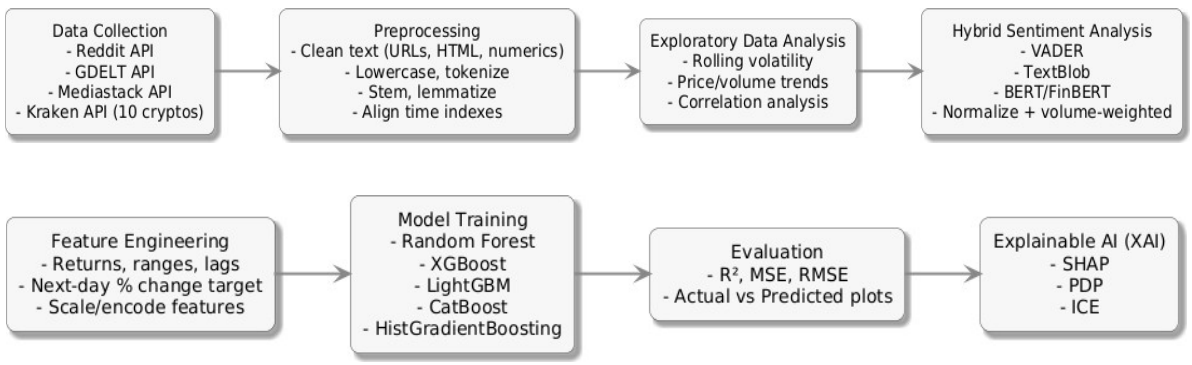

***Figure 1**: Methodology*

3.1 Data Collection

Using a variety of APIs, the dataset was gathered from numerous sources. The GDELT API (Global Database of Events, Language, and Tone) and the Mediastack API were used to collect Ethereum-related news headlines. While all cryptocurrencies' financial data came via the Kraken API, discussion forum data was acquired using the Reddit API. The diversity and depth of the dataset are increased by the news items from various geographical locations and sources that are provided by both GDELT and Mediastack. Discussions about cryptocurrencies' legal and regulatory issues are included in the Reddit data, which represents a variety of viewpoints and prejudices that influence new developments in the cryptocurrency market.

3.2 Data Preprocessing and Transformation

Data from three distinct domains was gathered and stored in ".csv" format, which included fields for the date, the news source or domain, and the associated URLs. The data, which covered the period from September 2024 to August 2025, revealed high-quality, real-time information and captured seasonal patterns. In accordance with [37], the text data was first cleaned by converting all the text to lowercase and eliminating URLs, HTML tags, and numeric characters to cut down on noise. A regular expression-based tokenizer was used to break up news headlines into individual words. Additional cleaning included eliminating common stop words like "and" "is," and "the," as well as words with fewer than three characters. This was done because

abbreviations frequently add noise and provide no relevant content when vetted. Words were reduced to their base forms using stemming and lemmatization, which is a common procedure in Natural Language Processing (NLP) workflows.
The open, high, low, and close prices of cryptos as well as the trading volume for each date were all included in the financial dataset. Based on their respective dates, this financial data and the news data were synchronized. The dataset had 11 characteristics and about 800 entries after merging. Following that, the consolidated dataset was ready for feature engineering, exploratory data analysis, and feature selection.

3.3 Feature Engineering and Exploratory Data Analysis (EDA)

Other features were designed and incorporated into the data to improve its quality and extract more significant insights.

Analysis of Price and Market Features cryptos financial data was examined to glean insights, and price fluctuations were computed using the formula below:

$$PriceChange = \frac{ClosePrice(USD) - LowPrice(USD)}{LowPrice(USD)} * 100$$

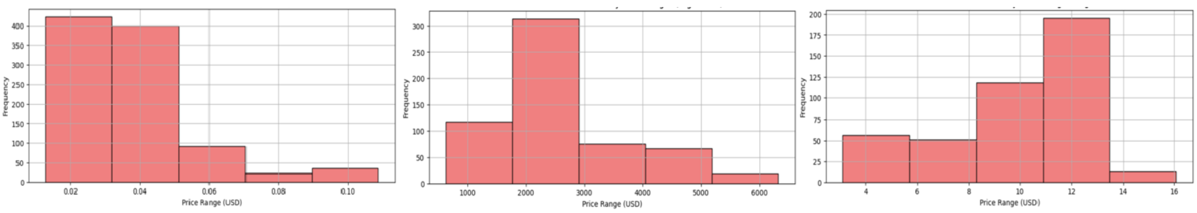

Figure 2. *Cardano, Bitcoin, Solana Distributed Price Ranges*
Both the deep learning (DL) and machine learning (ML) models used this feature as their target variable. Since closing prices frequently exhibit little variation from opening prices, forecasting price changes rather than Ethereum's closing price increases model reliability. Furthermore, focusing on price fluctuations reduces the danger of data leakage that could occur from using the opening price.
The use of same-day financial data (such as high, low, and volume), as shown in Fig. 2, as input features for forecasting same-day percentage price changes is a major worry. Such information would only be accessible following market close in a real-time forecasting scenario, which could result in data leakage. This creates uncertainty over the model's ability to operate as a real-time prediction system or as a post-hoc explanatory tool. Given this, the model would have to utilize just the features that are accessible at the forecast point to function in real-time. For instance, intraday data (partial volume, high, and low up to that point) and real-time mood data would be used to forecast end-of-day price moves at noon.
As an alternative, only the sentiment and closing data from the previous day would be used for next-day forecasts; future data would be excluded to avoid leakage. The model would be able to operate efficiently in a real-time or nearly real-time trading environment with these modifications.
Additionally, the high price was deducted from the low price to reflect market volatility and evaluate cryptos price stability.
The data was examined to determine the volatility factor during a seven-day period (Figs. 3 and 4). The visualizations of the number of investors demonstrate a close correlation with cryptos price.

Features of Sentimental Analysis In contrast to conventional financial assets, speculative public opinion, erratic social media trends, and political or geopolitical issues all have a significant impact on cryptocurrencies. Public opinion, feelings, and conversations surrounding the cryptocurrency market are greatly influenced by international news sources and websites like Reddit, which may have a direct or indirect effect on prices. Examining these sources provides important information on the psychology of the market.
The hybrid methodology used is intended to extract complex market sentiments from a variety of data sources.

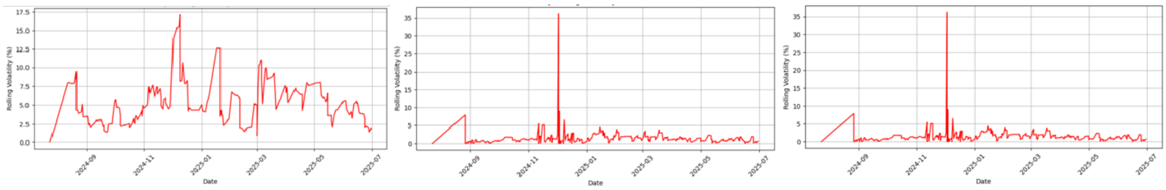

***Figure 3**. Chainlink, PolkaDOT, Litecoin Weekly Rolling Volatility*

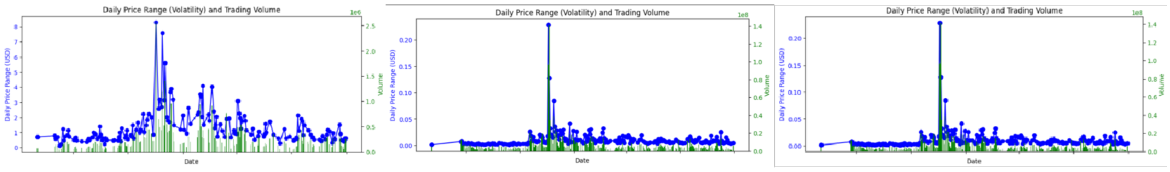

***Figure 4**. Chainlink, PolkaDOT, Litecoin Over Time Trading Volume*

A lexicon-based approach designed especially for social media sentiment analysis is called VADER (Valence Aware Dictionary and Sentiment Reasoner). Giving a quick and easy rule-based analysis, it uses a pre-defined lexicon of terms to assign sentiment scores between -4 and +4.
A transformer-based approach called BERT (Bidirectional Encoder Representations from Transformers) is popular for jobs involving natural language processing because it is excellent at comprehending sentence context. It captures nuanced and complicated meanings in the data while providing sentiment scores ranging from -2 to +2, which categorize text as positive, negative, or neutral.
Like VADER, TextBlob, a Python toolkit for NLP tasks, employs a lexicon-based methodology. By segmenting news headlines into smaller phrases and allocating scores using pre-trained lexicons, it determines polarity ratings ranging from -1 to +1.
The strengths of each of these sentiment analysis tools vary by domain. Their results differ according on the type of data, as Fig. 5 illustrates. A more precise relationship between market mood and fluctuations in bitcoin prices is established using the hybrid approach [38]. All three techniques were applied because to the delicate and nuanced nature of Reddit debates and news headlines. BERT can read more complicated settings, but it may also react to noise. In contrast, VADER and TextBlob are better at detecting strong sentiment signals.

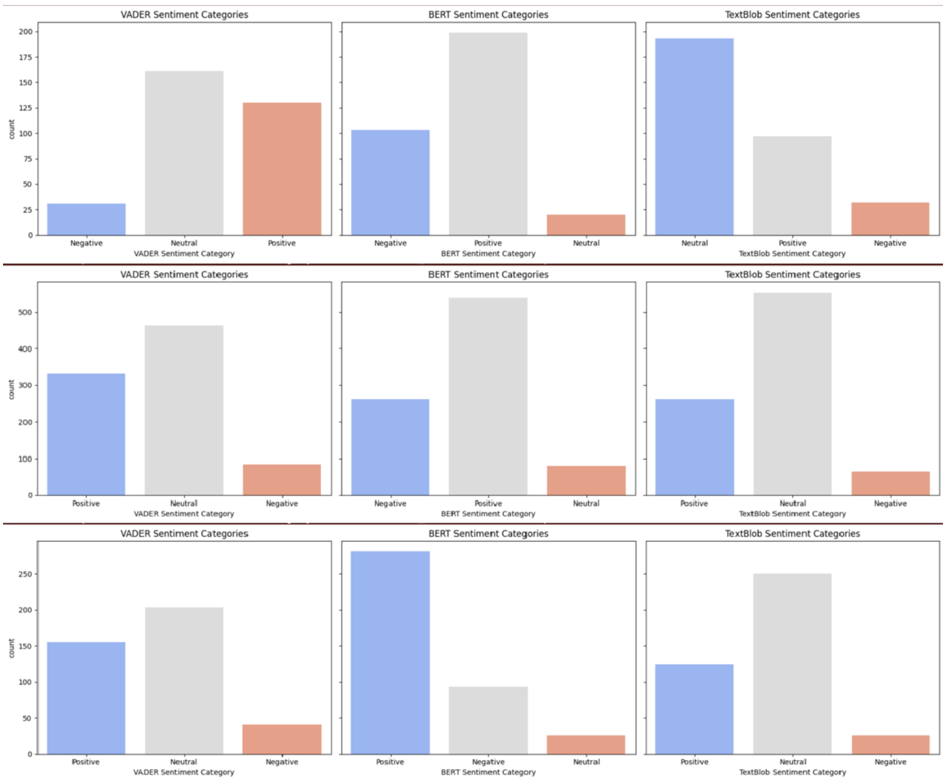

***Figure 5**. Solana, Tron, Chainlink Sentimental Analyser Visualizations*

Furthermore, not all Reddit content or financial news is specifically about finance, so sentiment scores may be impacted by subtle, non-financial language.
All scores were normalized to a [0, 1] range to standardize sentiment outputs; values close to 0 denoted negative sentiment, values close to 1 denoted positive sentiment, and mid-range values denoted neutrality. Min-Max scaling was used on all features to match sentiment ratings with financial indicators like price and volume, making sure that no feature was dominated by scale discrepancies. As a result, the model was able to discover harmonious connections between market behaviour and sentiment.
To combine the advantages of each approach, lower noise, and give the predictive model more dependable input, the average of the three sentiment ratings was finally calculated, as shown below.

$$AverageSentiment = \frac{VADERScore + BERTScore + TextBlobScore}{3}$$

It is acknowledged that the choice to mix the results of various sentiment analysers with volume and equally weight them is an empirically straightforward one. Equal weighting was selected for its ease of use, interpretability, and ability to lower the danger of overfitting given the small dataset, even if other aggregation techniques or learnt weightings might provide greater precision. A neutral, balanced aggregation of market sentiment is produced by equal weighting, and each sentiment analyser VADER, BERT, and TextBlob captures distinct nuances from the text. While more advanced techniques, such learning or performance-based weighting, may yield better outcomes, they come with greater complexity and demand larger datasets.
Investigating optimal weighting techniques is still a viable avenue for further study.
A correlation analysis between sentiment ratings and cryptos price fluctuations was carried out to better comprehend the connection between sentiment and market behaviour. The purpose of this investigation was to assist the model better capture this dynamic by reflecting how sentiment changes in response to price variations. Table 2,3,4 displays the correlation values for each sentiment score, their average, and the total hybrid sentiment scores.
In order to create a closer link between sentiment data and the market, cryptos trading volume—which is strongly correlated with price movements, as seen in Fig. 4 was integrated with the average sentiment scores.

Table 2,3: *Chainlink, Tron Correlation Analysis of the Sentimental Scores and Price Changes*

Sentiment Analysis	Price Changes Correlation
Hybrid Sentiments	0.704424
VADER	-0.012647
Average Sentiments	- 0.020921
BERT	- 0.068223
TextBlob	0.125401

Sentiment Analysis	Price Changes Correlation
Hybrid Sentiments	0.391481
VADER	0.482696
Average Sentiments	- 0.021701
BERT	- 0.043234
TextBlob	0.284219

This hybrid sentiment approach, which combines financial data and textual sentiment, produced a moderately positive correlation with price changes, indicating that it provides a more robust predictive signal for cryptos price movements than individual sentiment scores alone.
The hybrid sentiment approach overcomes the drawbacks of each sentiment analyser and has the following benefits:

- It provides a more thorough assessment of market sentiment by combining rule-based, lexicon-based (polarity), and context-based sentiment analysis.
- It improves the model's capacity to represent significant connections without superfluous complexity.
- By including mood data and trading volume, the research is grounded in the actual market.

The average hybrid sentiment scores are shown in Figs. 6 and 7, and the correlation analysis that goes with it acts as a gauge for the method's effectiveness.
Stationarity analysis was performed to confirm whether the statistical characteristics of cryptos time series data remained constant over time, ensuring reliable training of the machine learning and deep learning models. Stochastic and deterministic trends were assessed using the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests [39]. The ADF test produced a significantly negative statistic and

a p-value below 0.0002, as indicated in Table 3, supporting the null hypothesis' rejection and demonstrating stationarity.
Furthermore, demonstrating that no stationarity transformation was required, the KPSS statistic had a p-value above 0.05 and was significantly below all critical values.

Table 4. *Kwiatkowski- Phillips-Schmidt-Shin (**KPSS**) tests and Augmented Dickey-Fuller (**ADF**)*

ADF	KPSS
ADF Statistic = -4.4948	KPSS Statistic = 0.1082
p-value = 0.0002	p-value = 0.1
Critical Values 10% = -3.4389	Critical Values 1%: 0.739
Critical Values 5% = -2.8653	Critical Values 5%: 0.463
Critical Values 1% = -3.4389	Critical Values 10%: 0.347

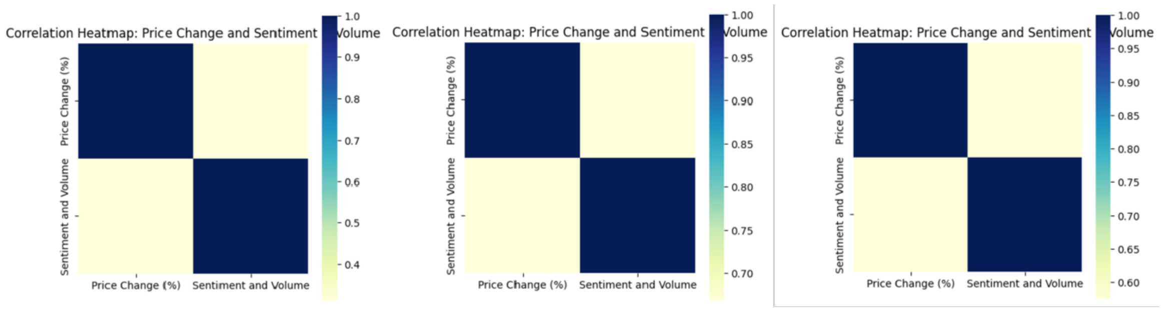

Figure 6. *Avalanche, Cardano & Ripple Correlation Analysis*

***Figure** 7. Avalanche, Cardano & Ripple Average Sentiments*

3.4 Modelling Approach and Evaluation Metrics

Several machine learning models were trained to predict daily percentage price changes for 10 cryptocurrencies: BTC, ETH, ADA, SOL, XRP, AVAX, LINK, DOT, LTC, and TRX. This was done after improving the quality of the data and designing important features. Using VADER, TextBlob, and FinBERT, the feature set integrated hybrid sentiment scores gleaned from Reddit discussions, GDELT news articles, and Mediastack headlines with financial indicators (OHLCV-derived metrics, volatility, and other technical features).
Tree-based ensemble models consistently outperformed other models in capturing complex and non-linear relationships between market data and sentiment, according to analysis. Strong predictive powers were shown by CatBoost, LightGBM, HistGradientBoosting, Random Forest, and XGBoost in particular; LightGBM and CatBoost regularly obtained the highest R^2 values across a variety of assets.
Tree-based approaches were resistant against the varied distributions across different cryptocurrencies because they handled heterogeneous feature types well and provided good interpretability through feature importance analysis. Both speed and accuracy were offered by LightGBM's histogram-based training and CatBoost's natural handling of categorical features. Random Forest and XGBoost continued to be dependable baselines for smaller datasets, whereas HistGradientBoosting provided a scalable scikit-learn-native substitute. The dataset focused on short-term market movements that were significantly impacted by recent mood and news, while covering a variety of assets at a daily resolution. The chosen tree-based techniques were picked for their stability, effectiveness, and performance under moderate sample sizes, even though deep learning techniques frequently need much bigger datasets to generalize well. To evaluate robustness across assets, held-out testing and cross-validation were used.

An accurate model should capture variance effectively, show low error, and provide interpretable results [34]. A good forecasting model should have minimal error, effectively capture volatility, and produce results that are easy to understand. Among the evaluation metrics were:

- **R^2:** Measures the proportion of variance explained by the model and its ability to capture complex relationships.
- **RMSE:** Reflects prediction accuracy in original units.
- **MSE:** Penalizes larger errors, providing a measure of model reliability.
- **MAE:** Measures average absolute error for robustness assessment.
- **Runtime:** Captures computational efficiency across models.

3.5 Explainable AI (XAI)

Model predictions were interpreted using XAI, with a focus on assessing how sentiment features affected cryptos prices [29], [40]. The top-performing models were subjected to feature significance, Individual Conditional Expectation (ICE), partial dependency plots (PDP), and SHAP. The efficacy of hybrid sentiment characteristics was validated using PDP and ICE, which demonstrated how they affected predictions.
Insights for risk assessment were obtained using SHAP analysis, which also demonstrated how individual traits, and their interconnections affected price volatility.
By making it clear how feature-engineered inputs affected results, these XAI strategies guaranteed the models' interpretability.

4 Algorithm and Design Implementation

By using feelings from Reddit discussion forums and international news headlines as crucial variables in forecasting cryptos price fluctuations, this study presents a novel hybrid technique. Algorithm 1 presents the algorithm's comprehensive workflow.

Algorithm 1 (Pseudocode):

1. Input: OHLCV financial data for BTC, ETH, ADA, SOL, XRP, AVAX, LINK, DOT, LTC, TRX;
    - news headlines from GDELT and Mediastack;
    - Reddit posts and comments.
2. Output: Hybrid sentiment scores for each asset, and their correlation with price changes.
3. for each cryptocurrency in asset list do
4. for each row in dataset do
5. Compute sentiment scores using:
    (a) VADER
    (b) TextBlob
    (c) FinBERT
6. Calculate the average of the three sentiment scores.
7. end for
8. Normalize average sentiment scores to [0, 1].
9. Hybrid Sentiment Score ← Weighted combination of normalized average sentiment and trading volume feature.
10. Compute correlation coefficients between:
    (a) Individual sentiment scores and price changes
    (b) Hybrid sentiment scores and price changes
11. end for
12. Compare predictive strength of hybrid sentiment scores against individual sentiment scores.

A MacBook Pro running macOS Ventura (Apple M1, 8 GB RAM) was used for the study. All calculations were carried out using Anaconda Navigator (Anaconda3) in a Jupyter Notebook environment using Python 3.10.2. Important Python tools used include Hugging Face Transformers (FinBERT), TextBlob, and VADER for sentiment analysis; Pandas and NumPy for handling and transforming data; Matplotlib and Seaborn for visualizing data; and Scikit-learn, LightGBM, CatBoost, HistGradientBoosting, RandomForest, and XGBoost for implementing and assessing machine learning models.

In order to integrate multi-source textual sentiment data with OHLCV-based financial measures for BTC, ETH, ADA, SOL, XRP, AVAX, LINK, DOT, LTC, and TRX, data gathering relies on APIs from Reddit, Mediastack, GDELT, and Kraken.

5 Implementation

Following extensive data cleaning, transformation, feature engineering, and exploratory analysis, the dataset was ready for model training. Text data from Reddit and international news was converted into multiple embeddings to enable comparative analysis based on the data's nature.

Four distinct categories of embeddings were employed:

1. TF-IDF is a simple yet powerful method for highlighting the importance of words. The 5000 top words were chosen, and thus high-dimensional sparse vectors were obtained.
2. Word2Vec - Employed 100-dimensional dense vectors to capture semantic similarity, with a window size of 5 and a minimum word count of 1.
3. BERT - Used the 'yiyanghkust/finbert-tone' model, which is fine-tuned for financial sentiment analysis, to capture sentence-level context.
4. SBERT - Utilized 'paraphrase-MiniLM-L6-v2' to produce optimized sentence representations, yielding high-quality context-aware embeddings.

BERT and SBERT, which were both pre-trained using transfer learning, were selected owing to their ability to understand complex, contextually varied text, especially useful for financial sentiment. Tree-based models, Random Forest, CatBoost, LightBGM, HistGradiantRegressor and XGBoost Regressors, were initially tested. Every model was trained individually on the various embeddings, with the data split 80/20 for training and testing, with a random seed of 42 for reproducibility. Random Forest used 100 estimators. XGBoost underwent hyperparameter tuning using Grid Search, trying trees (100, 150, 200), max depth (3, 5, 7), learning rates (0.01, 0.1, 0.2), and subsample/feature fractions (0.8, 1.0) and (0.3, 0.5, 0.8). Grid Search with 3-fold cross-validation guaranteed optimal parameter choice. Also, an LSTM model was trained on the same split and embeddings. It had 50 input nodes, 0.2 dropout rate for regularization, a ReLU the activation function and a single output neuron. The Adam optimizer was used for adaptive learning. The model worked optimally with 10 epochs and a batch size of 8. All models were critically compared, and the best-performing model received additional explainability analysis with XAI methods.

6 Evaluation

Regression assessment metrics, such as R2, RMSE, and MSE, as well as representations of actual versus predicted values, were used to compare the ML and DL models and determine how well each model captured variation and intricate relationships. Complexity of training time was also considered.

Table 6. *Random Forest Regressor model trained on different embeddings to predict the price changes for all cryptos.*

RANDOM FOREST REGRESSION				
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
BITCOIN				
TFIDF	0.000029	0.999266	0.005391	0.320729
Word2Vec	0.000047	0.998811	0.006862	0.695217
BERT	0.000136	0.996554	0.011682	5.390262
SBERT	0.000113	0.997143	0.010637	2.600984
ETHEREUM				
TFIDF	0.002120	0.940114	0.04604	1.298952
Word2Vec	0.003168	0.910525	0.056282	2.915239
BERT	0.004891	0.861846	0.069935	22.224222
SBERT	0.004673	0.867999	0.068360	11.393315
SOLANA				
TFIDF	0.000333	0.988567	0.018248	0.264064
Word2Vec	0.001743	0.940161	0.041746	0.703948

BERT	0.003969	0.863729	0.062997	5.296495
SBERT	0.003189	0.890514	0.056467	2.668276
CHAINLINK				
TFIDF	0.001869	0.769346	0.043226	0.452711
Word2Vec	0.003497	0.568316	0.059136	1.016422
BERT	0.003324	0.589623	0.057658	7.923267
SBERT	0.003146	0.611623	0.056091	3.874792
CARDANO				
TFIDF	0.000531	0.982804	0.023046	1.033838
Word2Vec	0.002619	0.915215	0.051174	2.367533
BERT	0.006063	0.803688	0.077868	19.259337
SBERT	0.005600	0.818697	0.074832	9.925278
TRON				
TFIDF	0.000332	0.971947	0.018222	1.436453
Word2Vec	0.000442	0.962630	0.021032	3.280080
BERT	0.000557	0.952923	0.023606	26.503709
SBERT	0.000571	0.951777	0.023891	13.678615
POLKADOT				
TFIDF	0.000272	0.991758	0.016502	0.808561
Word2Vec	0.004342	0.868580	0.065895	1.565420
BERT	0.010598	0.679234	0.102947	12.416418
SBERT	0.009962	0.698482	0.099810	6.375651
LITECOIN				
TFIDF	0.000422	0.973130	0.020554	0.961246
Word2Vec	0.000867	0.944866	0.029442	2.696054
BERT	0.001193	0.924143	0.034535	19.491194
SBERT	0.001118	0.928885	0.033438	10.273661
RIPPLE				
TFIDF	0.000297	0.953581	0.017229	1.135556
Word2Vec	0.000937	0.853500	0.030607	2.129135
BERT	0.001996	0.687899	0.044674	15.227839
SBERT	0.002301	0.640124	0.047972	7.582944
AVALANCHE				
TFIDF	0.000918	0.950941	0.030292	1.727261
Word2Vec	0.003034	0.837794	0.055080	2.891399
BERT	0.004726	0.747306	0.068748	22.656715
SBERT	0.004502	0.759283	0.067099	11.422085

The outcomes demonstrated how well the models performed in comparison to their design. Table 5 demonstrates how well the Random Forest Regressor captured the relationship between emotion and financial data, especially when TF-IDF embeddings were used. This implies that models employing context- aware embeddings are better able to capture deeper contextual information, whereas tree-based models such as Random Forest are better at utilizing the statistical importance of individual words.

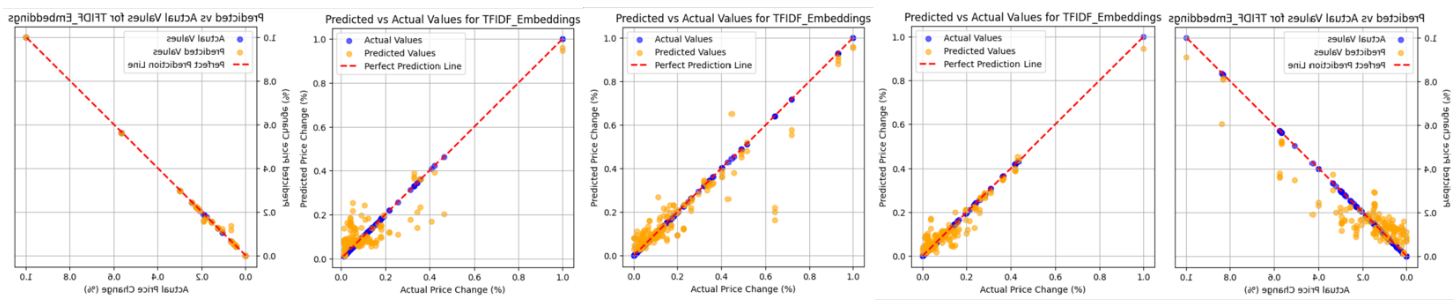

Figure 8. *Bitcoin, Avalanche, Ethereum, Litecoin & Ada RF Regressor Results for TF-IDF Embedding*

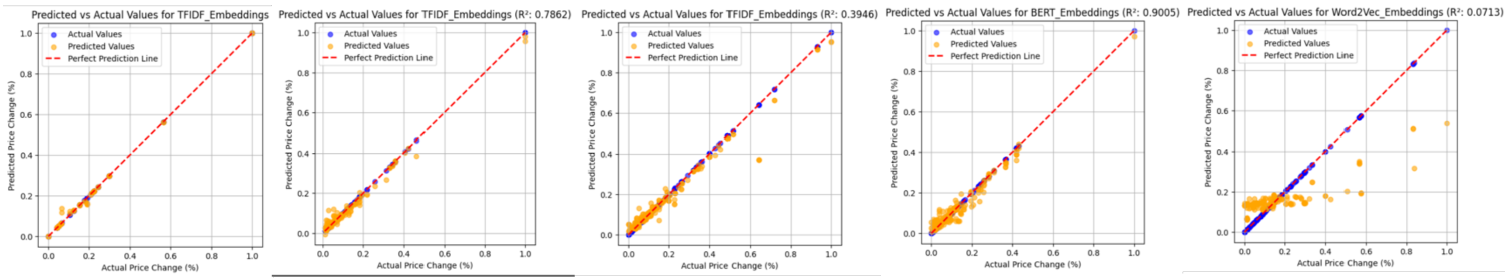

Figure 9. *Bitcoin, Avalanche, Ethereum, Litecoin & Ada XGB Regressor Results for TF-IDF Embedding*

By successfully capturing the variance in the dataset, the XGBoost model beat the other models; it did especially well with TF-IDF embeddings. XGBoost, a tree-based boosting algorithm, showed great ability to recognize both seasonal and short-term trends in the data. Like Random Forest, XGBoost demonstrated its ability to leverage word-level relevance over sentence-level context by performing better with embeddings that prioritize statistical correlations over contextual ones (as seen in Fig. 8,9 and Table 7).

Table 7. *XG BOOST Regressor model trained on different embeddings to predict the price changes for all cryptos.*

XG BOOST REGRESSOR				
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
BITCOIN				
TFIDF	0.001713	0.926238	0.041393	226.073672
Word2Vec	0.003183	0.862984	0.056415	34.801911
BERT	0.010907	0.530438	0.104437	125.391564
SBERT	0.006442	0.722681	0.08026	71.619131
ETHEREUM				
TFIDF	0.004167	0.394587	0.064553	281.293472
Word2Vec	0.011918	-0.731439	0.109168	39.533996
BERT	0.007736	- 0.123936	0.087955	142.987669
SBERT	0.007841	- 0.13918	0.08855	89.039229
SOLANA				
TFIDF	0.002058	0.644926	0.045369	154.715025
Word2Vec	0.009226	- 0.591523	0.096053	34.105707
BERT	0.001887	0.674412	0.043445	129.37423
SBERT	0.002341	0.596122	0.048387	70.458179
CHAINLINK				
TFIDF	0.000176	0.545253	0.013253	175.422767
Word2Vec	0.00109	- 1.821933	0.033016	29.806132
BERT	0.001385	- 2.585788	0.037217	115.358343
SBERT	0.000229	0.40666	0.015139	65.562747
CARDANO				
TFIDF	0.024484	-0.677471	0.156474	254.853238
Word2Vec	0.013555	0.071334	0.116425	45.101023
BERT	0.045543	-2.120272	0.213409	178.570234
SBERT	0.018935	- 0.297285	0.137605	110.334978
TRON				
TFIDF	0.000151	0.394268	0.012306	218.637262
Word2Vec	0.000338	- 0.351336	0.018381	38.734836
BERT	0.000272	- 0.088139	0.016494	146.823568
SBERT	0.00022	0.119213	0.01484	81.880742
POLKADOT				
TFIDF	0.012454	- 2.014762	0.1116	248.382456
Word2Vec	0.019967	- 3.833321	0.141305	37.337414
BERT	0.023373	- 4.657692	0.152882	135.208928
SBERT	0.020365	- 3.929522	0.142705	76.552301
LITECOIN				
TFIDF	0.003214	0.878637	0.056689	240.003074

Word2Vec	0.005014	0.810627	0.070813	61.945551
BERT	0.002634	0.900524	0.051323	226.393644
SBERT	0.003601	0.864026	0.060004	122.624283
RIPPLE				
TFIDF	0.001611	-0.1153	0.040142	319.973031
Word2Vec	0.001234	0.145812	0.03513	36.567349
BERT	0.001774	- 0.227952	0.042121	142.046323
SBERT	0.00149	- 0.031132	0.038598	79.731739
AVALANCHE				
TFIDF	0.001819	0.786205	0.042653	393.044989
Word2Vec	0.004765	0.440062	0.069027	40.51862
BERT	0.005843	0.3133	0.076442	150.751435
SBERT	0.003056	0.640879	0.05528	81.340767

CatBoost regressor performed really well with all the embedding but TFIDF gave a really promising results. Following table 8 showcased results for all the cryptocurrencies and embeddings. Like XG Boost and Random Forest Regression.

Table 8. *CAT BOOST Regressor model trained on different embeddings to predict the price changes for all cryptos.*

CAT BOOST REGRESSOR			
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error
BITCOIN			
TFIDF	4.42E-05	0.998912192	0.006648412
Word2Vec	0.001066064	0.973763883	0.03265064
BERT	0.001835007	0.954840005	0.042836981
SBERT	0.002201236	0.945827004	0.046917335
ETHEREUM			
TFIDF	0.001245367	0.961360111	0.035289761
Word2Vec	0.002026049	0.937137962	0.045011659
BERT	0.00369963	0.885211949	0.060824582
SBERT	0.003767217	0.883114925	0.061377659
SOLANA			
TFIDF	0.00239749	0.92607704	0.04896415
Word2Vec	0.00594846	0.81658818	0.07712625
BERT	0.00859255	0.73506152	0.09269602
SBERT	0.0060243	0.81424968	0.07761637
CHAINLINK			
TFIDF	0.002655226	0.05152889	0.6711333
Word2Vec	0.002960214	0.05440785	0.63335868
BERT	0.003718545	0.06097987	0.53943457
SBERT	0.003466036	0.05887305	0.57070936
CARDANO			
TFIDF	0.002189168	0.928022111	0.046788543
Word2Vec	0.003546221	0.8834034	0.059550157
BERT	0.007646652	0.748584879	0.087445137
SBERT	0.007240005	0.76195508	0.085088218
TRON			
TFIDF	0.00039069	0.9783599	0.01976588
Word2Vec	0.00139633	0.92265805	0.03736751
BERT	0.00221753	0.87717216	0.04709068
SBERT	0.00244284	0.86469261	0.04942507
POLKADOT			
TFIDF	0.00119327	0.96095865	0.0345437
Word2Vec	0.00437855	0.85674233	0.06617064
BERT	0.00680731	0.77727836	0.0825064
SBERT	0.00746217	0.75585258	0.08638384
LITECOIN			
TFIDF	0.00090932	0.94960591	0.03015494

Word2Vec	0.00154177	0.91455573	0.0392654
BERT	0.00254394	0.85901593	0.05043751
SBERT	0.00212297	0.88234597	0.0460757
RIPPLE			
TFIDF	0.00055166	0.92630833	0.02348746
Word2Vec	0.00101235	0.86476924	0.03181739
BERT	0.00144045	0.80758309	0.0379532
SBERT	0.00131757	0.82399648	0.0362984
AVALANCHE			
TFIDF	0.00182299	0.89113141	0.04269644
Word2Vec	0.00207357	0.87616676	0.04553643
BERT	0.00481522	0.71243553	0.06939178
SBERT	0.00613225	0.63378224	0.07830871

Like previous models, Light GBM regressor performed good with all the embedding but TFIDF gave a really promising results. Following table 9 showcased results for all the cryptocurrencies and embeddings. TFIDF is clearly the best of the embeddings.

Table 9. *Light BGM Regressor model trained on different embeddings to predict the price changes for all cryptos.*

LIGHT BGM REGRESSOR			
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error
BITCOIN			
TFIDF	0.000109224	0.997311976	0.010451016
Word2Vec	0.000140622	0.996539249	0.011858427
BERT	0.000299303	0.992634066	0.017300387
SBERT	0.000157973	0.996112249	0.012568722
ETHEREUM			
TFIDF	0.001490793	0.953745322	0.038610784
Word2Vec	0.002580055	0.91994888	0.050794243
BERT	0.004031162	0.874925537	0.063491431
SBERT	0.003551775	0.889799419	0.059596771
SOLANA			
TFIDF	0.00673971	0.79219103	0.08209575
Word2Vec	0.01000744	0.69143551	0.1000372
BERT	0.0127849	0.60579684	0.11307031
SBERT	0.0112481	0.65318171	0.10605705
CHAINLINK			
TFIDF	0.00330857	0.59021215	0.0575202
Word2Vec	0.0044832	0.44472699	0.06695671
BERT	0.00403848	0.49980793	0.06354907
SBERT	0.00354679	0.56070745	0.05955493
CARDANO			
TFIDF	0.002045537	0.93274457	0.04522761
Word2Vec	0.00617312	0.7970333	0.07856921
BERT	0.010300138	0.66134063	0.10148959
SBERT	0.009616593	0.68381498	0.09806423
TRON			
TFIDF	0.00027439	0.98480186	0.01656463
Word2Vec	0.00034344	0.9809768	0.01853226
BERT	0.00039392	0.97818093	0.01984745
SBERT	0.00042222	0.97661353	0.02054797
POLKADOT			
TFIDF	0.00057806	0.98108697	0.02404291
Word2Vec	0.00139228	0.95444723	0.03731331
BERT	0.00340517	0.88858969	0.0583538
SBERT	0.0031427	0.89717701	0.0560598
LITECOIN			
TFIDF	0.00043138	0.97609335	0.02076959

Word2Vec	0.00122801	0.93194434	0.03504295
BERT	0.00178846	0.90088453	0.04229016
SBERT	0.0022622	0.87462977	0.04756262
RIPPLE			
TFIDF	0.00050954	0.9319354	0.0225729
Word2Vec	0.00081093	0.89167476	0.02847683
BERT	0.00107132	0.85689083	0.03273109
SBERT	0.00121266	0.83801107	0.03482326
AVALANCHE			
TFIDF	0.00085909	0.94869549	0.02931016
Word2Vec	0.00230481	0.86235684	0.04800845
BERT	0.00669127	0.60039773	0.0818002
SBERT	0.00671612	0.59891369	0.08195195

Like previous models, Hist Gradient regressor performed good with all the embedding but TFIDF gave a really promising results. Following table 10 showcased results for all the cryptocurrencies and embeddings. TFIDF is clearly the best of the embeddings. Regression models worked well overall with TFIDF embedding.

Table 10. Hist Gradient *Regressor model trained on different embeddings to predict the price changes for all cryptos.*

HIST GRADIENT REGRESSOR			
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error
BITCOIN			
TFIDF	0.000140785	0.996444827	0.011865298
Word2Vec	0.000144521	0.996350499	0.012021677
BERT	0.000371681	0.990614143	0.019279034
SBERT	0.000201088	0.994922034	0.014180554
ETHEREUM			
TFIDF	0.000544527	0.984618893	0.023335112
Word2Vec	0.00125839	0.964454626	0.035473794
BERT	0.002915669	0.917641943	0.053996939
SBERT	0.002316827	0.934557281	0.048133425
SOLANA			
TFIDF	0.00159596	0.94519965	0.03994946
Word2Vec	0.00350314	0.87971302	0.0591873
BERT	0.00551206	0.81073276	0.07424324
SBERT	0.00567183	0.8052467	0.07531156
CHAINLINK			
TFIDF	0.00318485	0.60685618	0.0564345
Word2Vec	0.00345011	0.57411207	0.05873765
BERT	0.00339178	0.58131299	0.05823897
SBERT	0.00304505	0.62411331	0.055182
CARDANO			
TFIDF	0.00083939	0.97282376	0.02897218
Word2Vec	0.002865	0.90724179	0.05352573
BERT	0.00583616	0.81104674	0.07639477
SBERT	0.00629485	0.79619594	0.07934012
TRON			
TFIDF	0.0002429	0.97947906	0.0155852
Word2Vec	0.00033114	0.97202403	0.01819727
BERT	0.00039537	0.96659768	0.01988392
SBERT	0.00040476	0.96580414	0.02011873
POLKADOT			
TFIDF	0.0002593	0.99215188	0.01610279
Word2Vec	0.00111834	0.96615173	0.03344157
BERT	0.00382312	0.88428727	0.06183138
SBERT	0.00354571	0.89268345	0.05954588
LITECOIN			
TFIDF	0.00033282	0.97883174	0.01824327

Word2Vec	0.00100446	0.9361131	0.03169319
BERT	0.00166669	0.89399321	0.04082506
SBERT	0.00179079	0.88609964	0.04231775
RIPPLE			
TFIDF	0.00037473	0.94140034	0.01935784
Word2Vec	0.00069587	0.89118047	0.02637926
BERT	0.00095553	0.8505735	0.03091171
SBERT	0.001382376	0.7838239	0.03718032
AVALANCHE			
TFIDF	0.00068498	0.96337716	0.02617211
Word2Vec	0.00201772	0.89212133	0.04491904
BERT	0.00712079	0.61928254	0.08438478
SBERT	0.00622511	0.66717074	0.07889936

The evaluation of XGBoost only trained with hybrid sentiment features without any financial data with TF-IDF embedding showcases the independent prediction power of sentimental analysis in predicting crypto price changes. The R^2 scores are not reliable. They showcase moderate performance with accuracy ranging uptil ~50%. But, in a bigger image, it only tells us about the impact of the sentiments on the prices of the cryptocurrencies. Following table 11 will help understanding.

Table 11. *XG BOOST Regressor model trained on TF-IDF embeddings with only hybrid sentiments for all cryptos.*

XG BOOST REGRESSOR				
Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
BITCOIN				
TFIDF	0.053614848880580705	- 0.35390590872451333	0.23154880453282567	0.17
ETHEREUM				
TFIDF	0.038582802233307116	- 0.089837077498645	0.19642505500395593	0.27
SOLANA				
TFIDF	0.028505107957016336	0.021221793906026742	0.16883455794657778	0.14
CHAINLINK				
TFIDF	0.006173916838181616	0.23788075463758218	0.07857427593164074	0.16
CARDANO				
TFIDF	0.03655945346026182	- 0.18365970517535235	0.19120526525245538	0.39
TRON				
TFIDF	0.012029704099489013	- 0.01631318335363896	0.10968000774748793	0.19
POLKADOT				
TFIDF	0.03562846592085327	- 0.07835168243157087	0.18875504210710045	0.19
LITECOIN				
TFIDF	0.01807483368478303	- 0.1496198929461663	0.1344426780631174	0.21
RIPPLE				
TFIDF	0.008799217186199582	- 0.3760225092564302	0.09380414269209852	0.21
AVALANCHE				
TFIDF	0.017102094639081546	0.08562598374599273	0.13077497711367242	0.26

6.1 XAI Results

Explainable AI (XAI) approaches, such as Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots, were used to study the complexity of hybrid sentiment features and their interactions with other characteristics in order to evaluate their effect on model performance. These graphics made it easier to assess how well hybrid sentiment features and average sentiment scores predicted changes in the price of cryptos.
Although individual sentiment indicators were significant, the investigation revealed that hybrid sentiment features had a noticeably greater impact on the model's learning.

In comparison to average sentiment alone, hybrid sentiment features were able to capture a higher degree of non-linear variance in cryptos price fluctuations, as seen by the PDP and ICE plots (Figs. 10 and 11, respectively). This pattern suggests that the model gives hybrid sentiments greater weight, which enhances its capacity to more precisely forecast intricate market dynamics.

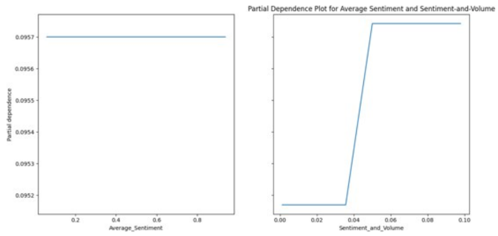

Figure 10. *PDP visualization of the Average and Hybrid sentiments for chainlink*

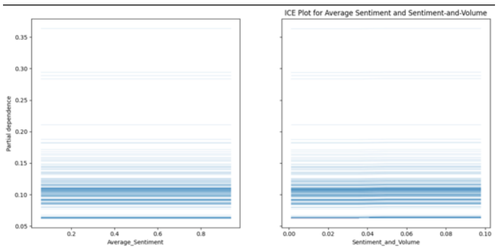

Figure 11. *ICE visualization for Average and Hybrid Sentiment for chainlink*

Using SHAP analysis, which was crucial in determining the efficacy of hybrid sentiment features during model training, the impact of these features on the first 25% of the test dataset was assessed. The findings demonstrated that hybrid sentiment features consistently and significantly improve predictions, with positive SHAP contributions in 39 out of 40 cases.

SHAP revealed the non-linear interactions between features and offered comprehensive insights into how each feature contributed to the model's output on an instance-by-instance basis. It emphasized the intricate dynamics that the model captures by highlighting the significance of each attribute separately as well as their interdependencies.
According to SHAP summary and interaction graphs (Figs. 12 and 13), trading volume and other financial indicators had a greater impact on predictions than sentiment elements. However, comparing hybrid sentiment features to individual and average sentiment scores was the main goal of the study. The results confirm that hybrid sentiment features enhance the model's learning process and increase the precision of crypto price forecasts, significantly improving model performance.

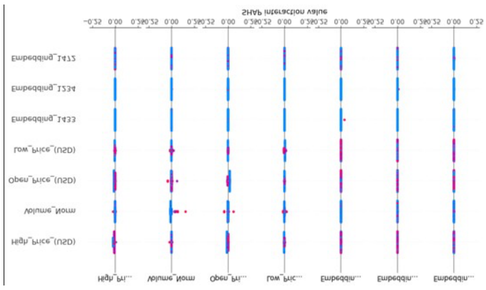

Figure 12. SHAP interaction value for tron

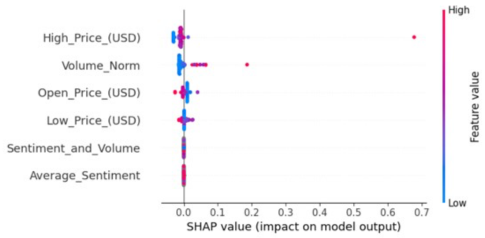

Figure 13. Impact of model's predictions using SHAP values for tron

The purpose of these assessments was to provide light on the interactions between features during model training and the relative contributions of each feature to the model's predictions.

6.2 Comprehensive results

Regression Algorithms worked well for almost all the datasets. The findings show that hybrid sentiment features provide a more comprehensive depiction of market sentiment by successfully reducing noise in text data. Sentiment analysis in model training was useful in showcasing machine learning and deep learning's potential for financial market prediction, especially when evaluating risk indicators for cryptocurrencies. This study's goal of examining investor psychology was accomplished when it found a significant correlation between sentiment data and fluctuations in crypto's price. The significance of combining sentiment with financial data is underscored by the necessity of extracting high-quality sentiment from complex text sources. All things considered, our study confirms that meticulous sentiment research greatly improves the precision and accuracy of crypto price forecasts.

7 Conclusion and Future Work

This study expanded its emphasis to ten key assets—BTC, ETH, ADA, SOL, XRP, AVAX, LINK, DOT, LTC, and TRX—selected for their liquidity, volatility, and market relevance considering cryptocurrencies' growing significance in the global financial ecosystem. In order to enhance short-term price forecasting, the study tackled the problem of combining market-based features with a variety of sentiment cues from various text sources.

By combining the outputs of VADER, TextBlob, and FinBERT, a hybrid sentiment framework was created and deployed to Reddit discussions and news headlines (using GDELT and Mediastack). To generate enriched feature sets, these sentiment ratings were standardized and integrated with volatility measurements, technical indicators, and trade volume. Rather of relying solely on individual sentiment indicators, this hybrid method sought to better capture complex investor mood and how it interacts with market fluctuations.

Using various embedding techniques, five tree-based ensemble algorithms—CatBoost, LightGBM, HistGradientBoosting, Random Forest, and XGBoost—were trained on all assets. Top performance was continuously attained by LightGBM and CatBoost, with XGBoost coming in second. The top-performing models were subjected to analyse feature contributions using explainable AI approaches such as Individual Conditional Expectation (ICE) curves, Partial Dependence Plots (PDP), and SHAP values. These results confirmed the importance of hybrid sentiment and trading volume in short-term cryptocurrency price dynamics by continuously highlighting their dominance in forming model predictions.

The results show that multi-source, high-quality sentiment data improves prediction performance for volatile digital assets, particularly when paired with trading activity. For traders, portfolio managers, and risk analysts looking for relevant short-term projections, this has real-world ramifications.

This study could be expanded in several ways by future research:

- Extending the data scope involves adding more sentiment channels, such Telegram, Twitter, or niche cryptocurrency communities.
- Temporal granularity: To capture ultra-short-term sentiment shocks, projections are being moved from a daily to an hourly or minute level.
- simulate diversity: Investigating hybrid ensembles that combine deep learning architectures with tree-based learners to simulate deeper feature interactions.
- Advances in XAI: Using sophisticated interpretability techniques to further demystify complex relationships, such as surrogate modelling or counterfactual analysis.
- Bias detection: Examining the influence of sentiment source biases on prediction validity, especially those originating from powerful investors or coordinated market behaviour.

By demonstrating the value of merging hybrid sentiment with technical indicators and enhancing transparency through XAI, our work offers a platform for more robust, interpretable, and actionable forecasting tools in cryptocurrency markets.

References

[39], D.A. and K. A. Ababio (2025) "Modeling Bitcoin Prices and Returns using ARIMA Model," 1, no. 3. Available at: https://ijid.kstu.edu.gh/index.php/ijid/article/view/22.

Ahmed, F.U. *et al.* (2023) "A Comparative Performance Evaluation of Bitcoin Price Prediction Using Machine Learning Techniques," in *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Dhaka, Bangladesh: IEEE, pp. 194–198. Available at: https://doi.org/10.1109/ICICT4SD59951.2023.10303490.

Anbaee Farimani, S., Jahan, M.V. and Milani Fard, A. (2024) “An Adaptive Multimodal Learning Model for Financial Market Price Prediction,” *IEEE Access*, 12, pp. 121846–121863. Available at: https://doi.org/10.1109/ACCESS.2024.3441029.

Angelini, M. *et al.* (2024) “A Visual Analytics Conceptual Framework for Explorable and Steerable Partial Dependence Analysis,” *IEEE Transactions on Visualization and Computer Graphics*, 30(8), pp. 4497–4513. Available at: https://doi.org/10.1109/TVCG.2023.3263739.

Armin, A., Shiri, A. and Bahrak, B. (2022) “Comparison of Machine Learning Methods for Cryptocurrency Price Prediction,” in *2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. *2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, Behshahr, Iran, Islamic Republic of: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/ICSPIS56952.2022.10043898.

Aslam, N. *et al.* (2022) “Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model,” *IEEE Access*, 10, pp. 39313–39324. Available at: https://doi.org/10.1109/ACCESS.2022.3165621.

Aslim, M.F. *et al.* (2023) “Utilization of LSTM (Long Short Term Memory) Based Sentiment Analysis for Stock Price Prediction,” *Asian Journal of Social and Humanities*, 1(12), pp. 1241–1255. Available at: https://doi.org/10.59888/ajosh.v1i12.141.

Baviskar, V.S., D., R. and Sankari, S.S.U. (2023) “Cryptocurrency Price Prediction and Analysis,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/ICCCNT56998.2023.10308332.

Bhadula, S., Kartik, D. and Gupta, D. (2024) “An Explainable AI Regression Model For Gold Price Prediction,” in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/I2CT61223.2024.10543640.

Cruz, L.F.S.A. and Silva, D.F. (2021) “Financial Time Series Forecasting Enriched with Textual Information,” in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA: IEEE, pp. 385–390. Available at: https://doi.org/10.1109/ICMLA52953.2021.00066.

Farimani, S.A., Jahan, M.V. and Milani Fard, A. (2022) “From Text Representation to Financial Market Prediction: A Literature Review,” *Information*, 13(10), p. 466. Available at: https://doi.org/10.3390/info13100466.

Fior, J., Cagliero, L. and Garza, P. (2022) “Leveraging Explainable AI to Support Cryptocurrency Investors,” *Future Internet*, 14(9), p. 251. Available at: https://doi.org/10.3390/fi14090251.

Gontyala, Sai Prasanna (no date) *Prediction of Cryptocurrency Price based on Sentiment Analysis and Machine Learning Approach*. National College or Ireland. Available at: https://norma.ncirl.ie/5163/.

Goodell, J.W. *et al.* (2023) “Explainable artificial intelligence modeling to forecast bitcoin prices,” *International Review of Financial Analysis*, 88, p. 102702. Available at: https://doi.org/10.1016/j.irfa.2023.102702.

Gupta, A. *et al.* (2023) “Cryptocurrency Prediction and Analysis between Supervised and Unsupervised Learning with XAI,” in *2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*. *2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, New Raipur, India: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/ICBDS58040.2023.10346583.

Jaiswal, S. *et al.* (2023) “Effect of News Headlines on Gold Price Prediction using NLP and Deep Learning,” in *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*. *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, Bangalore, India: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/ICAIA57370.2023.10169488.

Juyal, P. and Kundaliya, A. (2023) “A Comparative Study of Hybrid Deep Sentimental Analysis Learning Techniques with CNN and SVM,” in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*. *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, Sonbhadra, India: IEEE, pp. 596–600. Available at: https://doi.org/10.1109/AIC57670.2023.10263883.

Kamble, S., Muntean, C.H. and Simiscuka, A.A. (2024) “A Hybrid HSV and YCrCb OpenCV-based Skin Tone Recognition Mechanism for Makeup Recommender Systems,” in *2024 International Wireless Communications and Mobile Computing (IWCMC). 2024 International Wireless Communications and Mobile Computing (IWCMC)*, Ayia Napa, Cyprus: IEEE, pp. 1224–1229. Available at: https://doi.org/10.1109/IWCMC61514.2024.10592313.

Keya, A.J. *et al.* (2024) “Leveraging CatBoost and XAI for Enhanced Software Defect Prediction,” in *2024 27th International Conference on Computer and Information Technology (ICCIT). 2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox’s Bazar, Bangladesh: IEEE, pp. 2998–3003. Available at: https://doi.org/10.1109/ICCIT64611.2024.11021971.

Khandelwal, V., Varshney, H. and Munjal, G. (2024) “Sentiment Analysis based Stock Price Prediction using Machine Learning,” in *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT). 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Gharuan, India: IEEE, pp. 182–187. Available at: https://doi.org/10.1109/InCACCT61598.2024.10551038.

Malaichamy, G., Muntean, C.H. and Simiscuka, A.A. (2024) “Online Job Posting Authenticity Prediction with Machine and Deep Learning: Performance Comparison Between N-Gram and TF-IDF,” in A. Fred et al. (eds.) *Deep Learning Theory and Applications*. Cham: Springer Nature Switzerland (Communications in Computer and Information Science), pp. 143–162. Available at: https://doi.org/10.1007/978-3-031-66694-0_9.

Mercadier, M. *et al.* (2025) “Monitoring bank risk around the world using unsupervised learning,” *European Journal of Operational Research*, 324(2), pp. 590–615. Available at: https://doi.org/10.1016/j.ejor.2025.01.036.

Mercadier, M. and Lardy, J.-P. (2019) “Credit spread approximation and improvement using random forest regression,” *European Journal of Operational Research*, 277(1), pp. 351–365. Available at: https://doi.org/10.1016/j.ejor.2019.02.005.

Oikonomopoulos, S. *et al.* (2022) “Cryptocurrency Price Prediction using Social Media Sentiment Analysis,” in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA). 2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Corfu, Greece: IEEE, pp. 1–8. Available at: https://doi.org/10.1109/IISA56318.2022.9904351.

Parikh, A. *et al.* (2023) “Cryptocurrency Price Prediction using Regression Models on Momentum Indicators,” in *2023 6th International Conference on Advances in Science and Technology (ICAST). 2023 6th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India: IEEE, pp. 131–136. Available at: https://doi.org/10.1109/ICAST59062.2023.10455082.

Passalis, N. *et al.* (2022) “Multisource financial sentiment analysis for detecting Bitcoin price change indications using deep learning,” *Neural Computing and Applications*, 34(22), pp. 19441–19452. Available at: https://doi.org/10.1007/s00521-022-07509-6.

Raheman, A. *et al.* (2022) “Social Media Sentiment Analysis for Cryptocurrency Market Prediction.” arXiv. Available at: https://doi.org/10.48550/ARXIV.2204.10185.

Sahal, R. (no date) *Predicting Optimal Cryptocurrency using Social Media Sentimental Analysis*. National College of Ireland. Available at: https://norma.ncirl.ie/6643/.

Satish, Naresh Kumar (no date) *Forecasting Ethereum’s Price using ML and DL by Integrating Hybrid Sentiments in Multi-Source Market Data: Leveraging XAI*. National College of Ireland. Available at: https://norma.ncirl.ie/7965/.

Sexton, J.P. *et al.* (2021) “Automatic CNN-Based Enhancement of 360° Video Experience With Multisensorial Effects,” *IEEE Access*, 9, pp. 133156–133169. Available at: https://doi.org/10.1109/ACCESS.2021.3115701.

Sharma, K. and Bhalla, R. (2022) “Decision Support Machine- A Hybrid Model for Sentiment Analysis of News Headlines of Stock Market,” *International journal of electrical and computer engineering systems*, 13(9), pp. 791–798. Available at: https://doi.org/10.32985/ijeces.13.9.7.

Simiscuka, A.A., Yaqoob, A. and Muntean, G.-M. (2024) “Utility-Based Multipath Delivery of Prioritized XR Content in a Machine Learning and Network Slicing-enhanced Environment,” in *2024 International Wireless Communications and Mobile Computing (IWCMC). 2024 International Wireless Communications and Mobile Computing (IWCMC)*, Ayia Napa, Cyprus: IEEE, pp. 1637–1642. Available at: https://doi.org/10.1109/IWCMC61514.2024.10592350.

Singh, M. *et al.* (2023) "Machine Learning Based Framework for Cryptocurrency Price Prediction," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*. *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India: IEEE, pp. 31–36. Available at: https://doi.org/10.1109/ICSCCC58608.2023.10176572.

Swati, S. and Mohan, A. (2022) "Cryptocurrency Value Prediction with Boosting Models," in *2022 International Conference on Intelligent Innovations in Engineering and Technology (ICIIET)*. *2022 International Conference on Intelligent Innovations in Engineering and Technology (ICIIET)*, Coimbatore, India: IEEE, pp. 183–188. Available at: https://doi.org/10.1109/ICIIET55458.2022.9967540.

Szabó, P. *et al.* (2023) "A CNN-Based Framework for Enhancing 360° VR Experiences With Multisensorial Effects," *IEEE Transactions on Multimedia*, 25, pp. 3245–3258. Available at: https://doi.org/10.1109/TMM.2022.3157556.

Tahmid Akhand, Md.N., Ahsan Habib, Md. and Rokibul Alam, K.Md. (2023) "Analyzing Cryptocurrency Price Trends for Real-Time Price Predictions," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/ICCIT60459.2023.10441450.

Ye, X. *et al.* (2022) "A Crypto Market Forecasting Method Based on Catboost Model and Bigdata," in *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*. *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Xi'an, China: IEEE, pp. 686–689. Available at: https://doi.org/10.1109/ICSP54964.2022.9778789.

Yu, D. (2022) "Cryptocurrency Price Prediction Based on Long-Term and Short-Term Integrated Learning," in *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*. *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, Shenyang, China: IEEE, pp. 543–548. Available at: https://doi.org/10.1109/ICPECA53709.2022.9718963.

Zhao, X., Liu, Y. and Zhao, Q. (2023) "Cost Harmonization LightGBM-Based Stock Market Prediction," *IEEE Access*, 11, pp. 105009–105026. Available at: https://doi.org/10.1109/ACCESS.2023.3318478.

Zubair, M. *et al.* (2024) "An Improved Machine Learning-Driven Framework for Cryptocurrencies Price Prediction With Sentimental Cautioning," *IEEE Access*, 12, pp. 51395–51418. Available at: https://doi.org/10.1109/ACCESS.2024.3367129.