Forecasting Ethereum's Price using ML and DL by Integrating Hybrid Sentiments in Multi-Source Market Data: Leveraging XAI

MSc Research Project
Artificial Intelligence

Naresh Kumar Satish
Student ID: 23248441

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Naresh Kumar Satish
Student ID:	23248441
Programme:	Artificial Intelligence
Year:	2024-25
Module:	MSc Research Project
Supervisor:	Anderson Simiscuka
Submission Due Date:	12/12/2024
Project Title:	Forecasting Ethereum's Price using ML and DL by Integrating Hybrid Sentiments in Multi-Source Market Data: Leveraging XAI
Word Count:	7389
Page Count:	24

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:	Naresh Kumar Satish
Date:	25th January 2025

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 Ethereum's Price using ML and DL by Integrating Hybrid Sentiments in Multi-Source Market Data: Leveraging XAI

Naresh Kumar Satish
23248441

Abstract

The cryptocurrency market, which is considered to be one of the most volatile markets due to its inconsistencies in the pricing factors, is yet widely used by a larger population incurring losses in most of the cases. To act as a risk assessment factor among the investors, users and other groups, the research study leverages the concept of forecasting Ethereum's prices by analyzing its social media sentiments like global news headlines, Reddit discussion forums and enhancing the data with hybrid sentimental features derived from VADER, BERT, Text Blob and correlating them with the financial parameters of the Ethereum to build a strong relationship among them and train machine learning models. The study has showcased prediction results of Ethereum using Random Forest, Extreme Gradient Boosting and Long Short-Term Memory models, critically evaluated for various factors and visualized that Extreme Gradient Boosting outperforms the other two models in capturing the complex relationship in the data and presenting a R-squared value of 0.982115. The study has presented the critical evaluation of the models, justification of the model's results and limitations. To enhance the risk assessment application of the study, the concept of explainable AI has been utilized to have transparency and accountability in the model's results. Shapley Additive Explanations (SHAP) is incorporated in the research study to explain the XG Boost's model's interaction on the features enhancing the reliability of the model. And concluding with some of the limitations of the model's performance regarding the nature of data.

1 Introduction

Cryptocurrencies have gained huge importance in society reshaping the financial markets due to the properties and applications of decentralized systems, financial freedom and enabling an independent global financial society. Currencies such as Bitcoin and Ethereum, an open-source network, have gained a lot of popularity, importance and have always been trending in the markets for the past few decades. The pricing factors are so volatile that it keeps the cryptocurrency's market unstable loosing trust among the investors. The research study utilizes Ethereum to study its influencing factors, volatility parameters and aims in building a strong and stable market which will be trustworthy among the investors, society, financial analysts, and institutions. Market analysis proves that the pricing factors are strongly dependent on the volume of the investors, and it is common known factor that the investors are influenced by the global news about the cryptocurrency and discussions about it. Hence the research work chooses this data to correlate with the activity of the investors and train the machine learning and deep learning models with the integration of both.

With the recent innovations and advancements in the field of machine learning, deep learning, natural language processing (NLP) and image processing, researchers work on maximizing the applications of these technologies to all the sufficient applications in the market, the research study focuses on analyzing the strong relationship and the complex patterns involved in the influencing factors of Ethereum like the social media posts, global news, financial parameters (opening price, closing price, low price, volume of the investors) and aims in forecasting the price of Ethereum in the latest market. Correlating the utmost parameters of Ethereum and bringing in a strong relationship within them to align with the volatile prices using machine learning (ML) and deep learning (DL) has been the major contribution of the study. While previous research studies have shown the sentimental analysis in predicting the stock markets and cryptocurrency prices by not critically assessing the financial data parameters. **To the best of our knowledge this is the first research work that combines hybrid sentimental analysis of news articles correlating them to financial parameters in order to predict prices of Ethereum.** This contribution is the main novelty of the study and leads to the research question,

"How effective are Random Forest, Extreme Gradient Boosting and Long Short Term Memory (LSTM) models in forecasting Ethereum's prices based on the hybrid sentimental scores of Ethereum's global news headlines, Reddit discussion forums and its financial data?"

"How can explainable AI algorithms(SHAP) demonstrate the feature interactions on the best working model, achieving increased transparency?"

The objective of the research work is to strongly correlate the market data with the activity of the investors so that there is influence of the market sentiments on the pricing factors of the Ethereum cryptocurrency. Contributing the research highlights, the importance of social media sentiments combined with other financial parameters acts as a major factor in predicting Ethereum's price by analyzing its volatility in the market. The importance of covering the sentiments of the subtle social media using different sentimental analyzers is evident by implementing machine learning models like Random Forest, Extreme Gradient Boosting and Long Short-Term Memory to predict the prices. The sections in the research paper are addressed as follows: 2 is the related work analysis, 3 highlights the overall methodology involved in the work, 4 outlines the algorithm and design implemented, 5 states the implementation of ML and DL models and the metrics used for evaluation, 6 and 7 concludes the research by highlighting the results, discussions and conclusion about the work.

2 Related Work

This section highlights the critical analysis of the recent research studies conducted relevant to the current research work. The published previous research studies are evaluated for their novelty in the work, methodology, quality of the dataset, algorithms implemented, evaluation metrics used and aims in building a forecasting model which addresses their limitations. Depending on different phases of the implementation, the analysis is divided into subcategories;

2.1 Sentimental Analysis

The collected data involves news headlines and reddit discussion forums(text), the research work aims to analyze the sentiments in the data and to correlate them with the prices, the research studies are analyzed for their implementation of sentimental analysis on various sources of text data which lead to effective forecasting of prices.
(Zubair et al., 2024) has showcased the hybrid model strategy in detecting cryptocurrency price by using real time data from twitter, CoinDesk and CoinMarketCap using API's. The study has developed a hybrid model by combining Bi-LSTM and GRU and has compared them among various other deep learning and machine learning models. The text data was analyzed by combining the scores from VADER and BERT to enhance the reliability of the prediction models. The hybrid model showcases an average MSE for Bitcoin – 0.024, Ethereum – 0.064, Dogecoin – 0.097.

Likewise, (Sharma and Bhalla, 2022) has proposed another hybrid model which uses Decision Tree and Support Vector Machine in classifying the sentimental groups as positive, neutral and negative. The study uses financial news headlines from various IT companies namely, TCS, Tech Mahindra, Coforge, HCL, Infosys and Wipro and their corresponding stock price data from the NSE website. The aim of the research is to compare various models in finding out which models show highest accuracy in classifying the sentimental groups accordingly. Initially the decision tree model is implemented to classify the groups which mainly focuses on classification of positive groups, SVM handles text data which is termed to have some subtle contexts and classifies them more accurately. Hyperparameter optimization of the hybrid model is done using Tune – sklearn package by fine tuning the model's kernel and regularization parameters. By this method the research work showcases that hybrid model is used to classify the sentiments accurately achieving an overall accuracy of 0.7975 which is comparatively very high when compared to other models in classification.

Another time series price prediction of cryptocurrencies was performed by (Oikonomopoulos et al., 2022) which utilizes twitter data and expresses the volatility of the cryptocurrency by the tweets posted by certain accounts. The study also uses VADER as their sentiment analyzer tool and performs sentiment analysis on twitter data. Finally, the sentiment analysis of VADER integrated with the transaction data of the cryptocurrencies, serves as the data for training time series model – Vector Autoregression, which showcases an accuracy score of 99.67.

The research work by (Aslim et al., 2023) explores the application of sentimental based stock market price prediction by calculating the sentimental scores using TextBlob sentiment analysis method on the news data, which is retrieved from CNBC Indonesia's news portal, the study uses sentimental scores as an extra factor for enhancing the price prediction along with the stock details. The results were evaluated for predictions which did not include sentimental scores and predictions which included sentimental analysis of different methods, lexicon based and TextBlob analysis. The predictions without sentiments arrived at a result that sentimental scores have a marginal impact on the price prediction and the study puts forwards many limitations due to this model's results.

The evaluation of hybrid methodologies continues by another research study by (Cruz and Silva, 2021), proposing the applications of BERT model in analyzing the sentiments of the textual data, the data includes. The dataset comprised of news headlines collected from the financial websites, complications in retrieving public source data are addressed and the data is collected accordingly. The study uses comparisons of various models in analyzing the data sentimentally and forecasting the prices, it was evident that DistilBERT model showed accuracy of 98.00 in analyzing the sentiments in the textual data and outperformed all the other models. Similarly, another research study by (Passalis et al., 2022) also showcased the applications of BERT model in analyzing the sentimentally on a similar type of data but only varying for its sources. The study compares the analysis of various kinds of BERT models, and the results show cased that CryptoBERT model showed an overall accuracy of 0.92 outperforming other BERT models. The challenges faced in the sentimental analysis was dealing different kinds of data and using different models depending on the nature of the data and applications of the models.

2.2 Explainable AI for cryptocurrency

The importance of explainable AI methods has been considered as a much-needed aspect in today's world, since from the exploration of complex deep learning models and tree-based algorithms, it has always been a question of model's transparency and accountability. The referred works were analyzed for explainable AI techniques in the same domain – cryptocurrency and financial markets. To address the volatility of one of the most trending and highly priced cryptocurrency – Ethereum, explainability techniques were used in the prediction of prices for bitcoin by (Goodell et al., 2023). The study explores the dataset features to explore about their feature importances rankings. It was mainly addressed to witness the volatility of Bitcoins during a political and geographical war (Russia-Ukraine war). SHAP algorithm was implemented on machine learning models such as Gradient Boosting and Random Forest to focus on the price determining factors, the values provided by the algorithm were able to derive at a conclusion on the factors which had most impact on the price predictions of Bitcoin.

(Gupta et al., 2023) does a comparative analysis between supervised and unsupervised machine learning in predicting the prices of the cryptocurrencies, the dataset used for the prediction analysis includes only the details of the cryptocurrencies like open price, low price, closing price and other relevant ones. The supervised models implemented in the forecasting are Random Forest, Support Vector Machine and the unsupervised models are LSTM and GRU. The study explores the use of SHAP algorithm in explaining each of the model's decision in predicting the prices, and it was identified that features like Open price, High price of the day, and low price of the day were identified has less significance in the model's decisions. The comparative analysis makes the work unique in explaining different weightages assigned to the features of different algorithms.

Another study by(Raheman et al., 2024) was about using XAI algorithms on models which used sentimental analysis for cryptocurrency market prediction, the transparency was explained for model such as Aigents with n- grams, the research focused on visualizing the XAI results of the model to have better explanatory in the conclusion. Visualizations of sentimental scores and their respective correlation with the cryptocurrency prices were implemented.

Leveraging XAI for gaining trust in the investors was the main aim of the research study proposed by (Fior et al., 2022), which is also similar work to other related studies, but study employed dataset which uses 21 different cryptocurrencies and was technically analyzed for its variability over various seasons, events, and other factors. SHAP scores explained the model's weightage criteria in the background providing necessary technical insights to the investors to focus on their investments valuing all the important factors that influenced the market.

2.3 Text Representation

Representing the text data into embedding has many advantages and disadvantages from capturing the contextual information of the news headlines to frequency of the words occurring in the document. Analyzing the research studies with the same context of predicting the prices of either cryptocurrency and financial market's news headlines, twitter data, and other social media. (Jaiswal et al., 2023) analyses the impact of news headlines in predicting the price of the gold, the author collects the general news headlines data, and uses TF-IDF, GloVe, and BERT for representing the texts into embeddings and the results showed that, the models which trained with BERT embeddings have shown good and accurate results. Since BERT embeddings were able to capture the contextual meaning within the large dataset of news headlines.

Another research which arrives at a result where BERT model again outperforms the other traditional embeddings, (Farimani et al., 2024) used FinBERT model for representing the texts into embeddings and trained Recurrent Convolutional Neural Network for predicting the price series in financial markets and the results outperformed the current state of the art models. Specifically, the authors have used concept-based representation – BERT BoEC (Bag of Economic Concepts) which is method to cluster the topics semantically. The advantage of this concept works on large dataset which is derived from a similar domain, and it also helps in capturing subtle contextual meaning from the news headlines.

It depends on the size of the dataset, quality of the dataset and many other factors which have impact on the text embeddings of different algorithms. To showcase an example of this, (Gontyala, 2021) has used news headlines which were collected between a short time span and that is relatively small dataset when compared to others. And the research work was evident in showing that on smaller datasets, TF-IDF vectorization techniques were much more advantageous that other large models like BERT and FinBERT. LSTM model was trained on these embeddings to predict the prices of the Bitcoin, and the model was fine tuned to obtain good results.
(Sahal, 2022), (Rateb et al., 2024) and (Farimani et al., 2022) have also demonstrated their research studies using same type of dataset – twitter data and google trends. The authors have used this dataset to predict the price of the cryptocurrency and have also shown that methods like BERT, GloVe and TF-IDF vectorization techniques have been showing good results in enhancing the prediction of the cryptocurrency prices.

2.4 Machine learning and deep learning algorithms for stock and crypto price predictions

To critically evaluate on the list of machine learning and deep learning algorithms, several research studies with similar aim have been analyzed and summarized in the below table,

Authors	Dataset	Research Aim	Model Implementation	Model Evaluation
(Munjal et al., 2024)	Twitter data and ICICI bank stocks	Stock price prediction	kNN, SVM, LR and RF	LR - 0.8741
(Bhadula et al., 2024)	Yahoo finance data	Gold price prediction	kNN, RF, MLP, HR, SVR, GBR, AdaBoost and CatBoost	kNN - 0.9999957
(Singh et al., 2023)	Coinmarketcap data	ETH price prediction	ARIMA, FbProphet and Trees	ARIMA-produced less errors
(Akhand et al., 2023)	Real time crypto data - twelvedata.com	Crypto price prediction	CNN-LSTM, CNN-GRU, and CNN-BiLSTM	CNN-LSTM - RMSE of 235.97,
(Armin et al., 2022)	Yahoo finance data	BTC, ETH, BNB price prediction	LR, RR, DNN, RNN, RF, kNN, LSTM	Ridge - MSE of 0.000202
(Yu, 2022)	Crpytocurrency websites data	Crytpo price prediction	SVR - integrated learning	0.783015776

Table 1: Summary of the ML and DL algorithms implemented

the table summarizes the dataset used, algorithm implemented, aim of the research work along with the model's results. The limitations of the research studies are addressed to be data quality factors, use of hybrid models, overfitting of the models and lack of contextual information in the news headlines due to which, the model struggles to correlate the prices with the sentiments and embeddings.

2.5 Discussions about Related Work

From the related work analysis, the various challenges were addressed regarding the sentimental analysis of the text data like dealing with the subtleness in the dataset, the macro-indicators present in the texts which is mainly domain specific can act as a noise to the ML and DL models but may carry contexts, and the biases in the sentiments especially when dealing with the unlabeled data, XAI methods implemented, text representations depending upon the nature of the data and the model, and limitations of the machine learning and deep learning models. Briefly, the analysis was able to help me by addressing that sentimental factors involved in the social media that vary seasonally depending on the seasonal trends and events, hence it will be accurate to have comparative analysis or hybrid creation to bring the best quality of sentiments in the data, XAI methods were adopted based on their applications in the model's feature interactions and the overall contribution in the model's predictions, Text representation has its application over different nature of the data and different applications of the model over the course. Hence the research work aims in addressing all these challenges, applications and trends involved.

3 Methodology

The research work aims at predicting the Ethereum's price changes by integrating the global news headlines data, Reddit discussion forums data and financial data of the Ethereum cryptocurrency. The work implemented is based on analyzing the relationships between the financial data and the sentiments, following a feature engineered hybrid sentiment and financial analysis pipeline for predictive modeling methodology. Furthermore, the research study explains the model's transparency in analyzing the feature interactions and contribution involved in the model's predictions additionally serving as an evaluation metrics to assess the hybrid sentimental feature and explaining the models transparency using Shap algorithm. Fig 1 shows the basic research methodology and the workflow executed in the research study.

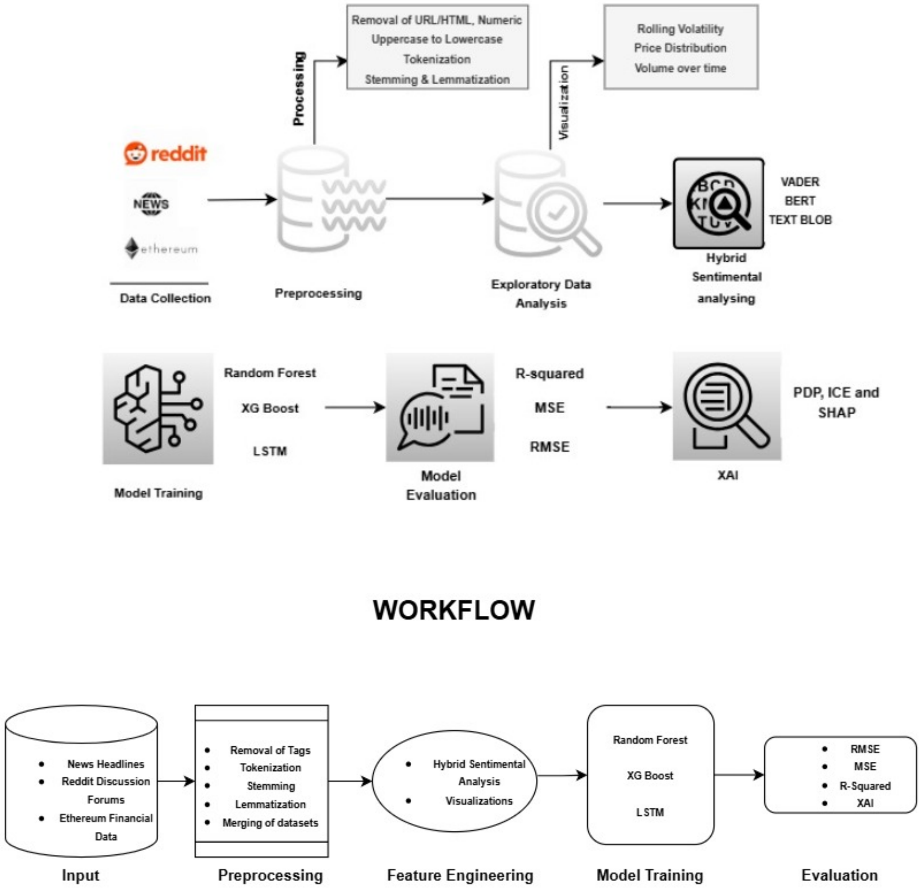

Figure 1: Research Methodology

3.1 Data Collection

The dataset was retrieved and collected using various APIs from multiple sources, the news headlines regarding the Ethereum cryptocurrency was retrieved using GDELT API (Global Database of Events, Language, and Tone), Mediastack API, and the data from the discussion forums was retrieved using Reddit API. And Ethereum's financial data was sourced using Kraken API. Mediastack and GDELT sources contain news articles of various sources, across diverse locations enhancing the quality of the dataset. Reddit data comprises of the discussions of the cryptocurrency's legal information, and which tends to have various opinions and biases creating trends in the cryptocurrency market.

3.2 Data Processing and Transformation

The data from 3 different domains were sourced in a ".csv" format with the date, source/domain of the news, and its URL. The time span of the data retrieved was from May 2024 to October 2024 to represent seasonal trends, real time and high-quality data. Referencing the work of (Aslam et al., 2022), the text dataset was initially cleaned by removing the URL/HTML tags, removal of numeric characters to eliminate the noises in the data, and the data was converted into lowercase. The sentences in the news headlines were split into separate words and was tokenized using tokenizer based on the regular expressions, the data was further cleaned by removing the stop words like ‘and’, ‘is’, ‘the’ and the data was filtered. And it is common that news headlines are more likely to comprise of abbreviations which serves as an outlier to the model bringing no context during vectorization, hence the words which had fewer than three characters were eliminated. And the words were reduced to their base form using stemming and lemmatization which is a common preprocessing step in Natural Language Processing (NLP).

The financial data includes the Ethereum’s date, open price, high price, low price, close price, and volume on the respective dates was analyzed and it was merged with the Ethereum’s news data according to their date. The dataset after merging had close to 800 data values with 11 features. The merged dataset was prepared for further feature engineering, exploratory data analysis and feature selection.

3.3 Feature Engineering and Exploratory Data Analysis

To enhance the quality of the dataset and bring meaningful insights, the data was engineered with various additions of new features.

3.3.1 Price and Market Features Analysis

The financial data of Ethereum was analyzed to obtain insights and its price differences/changes of Ethereum was calculated using,

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

This feature serves as the target variable for the ML and DL models implemented, predicting this feature instead of the closing price of Ethereum will make the model more reliable. Since some prices in the closing data tend to show minute deviation from the opening price. There could be also probability that opening price could act as a data leakage factor, hence price changes are set as the target variable. To represent the volatility of the Ethereum cryptocurrency in the market and to visualize the market stability, its high price was differenced with its low price,

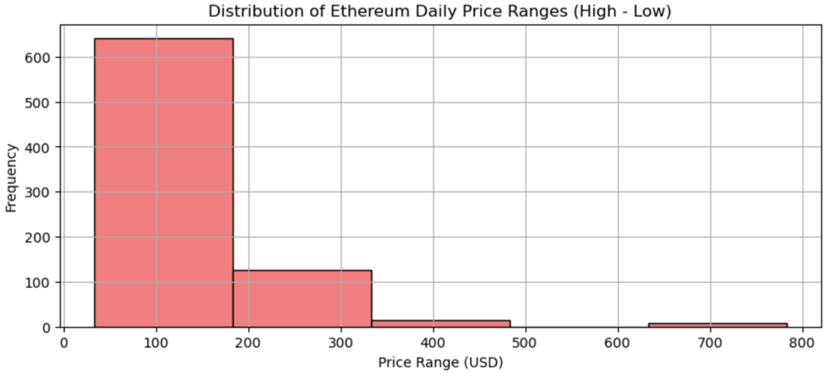

Figure 2: Distribution of Ethereum Price ranges

The data was analyzed for showing the volatility factor within a time span of 7 days, which could be another factor in questioning the model’s capability to predict unstable data,

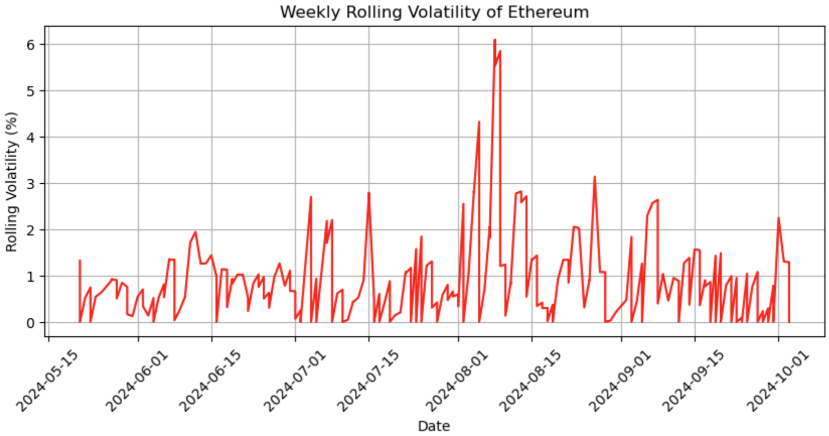

Figure 3: Weekly Rolling Volatility of Ethereum

The visualizations over the volume of investors show a strong relationship with the price movements of Ethereum,

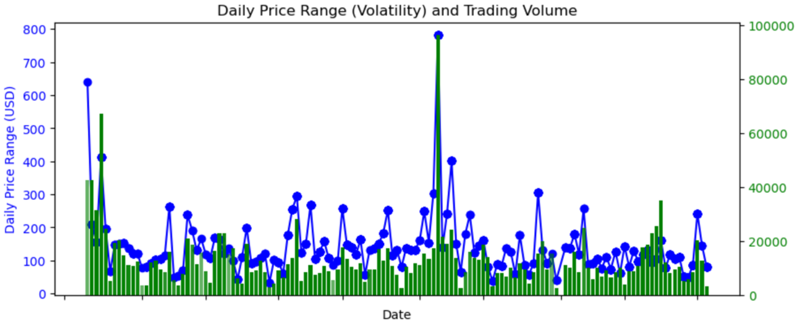

Figure 4: Ethereum Trading Volume Over Time

3.3.2 Sentimental Analysis Features

Unlike other financial assets, cryptocurrencies are influenced by various speculative behavior of public sentiments, volatile trends in the social media, other political and geopolitical behaviors. Global news and platforms like Reddit serve as major platforms for the public users to convey their thoughts, emotions and also discussions about the cryptocurrency market which triggers the currencies either directly or indirectly. Analysing these will significantly be an attentive measure of the market psychology. The implemented methodology is a hybrid method in analyzing the market sentiments from nuanced data.

VADER (Valence Aware Dictionary and Sentiment Reasoner), a lexicon-based approach, more specifically designed for analyzing the sentiments of the social media. The approach is basically characterized by pre-defined words, each associated with a sentiment score ranging from -4 to $+4$, this gives the sentiment intensity of the words. It is a faster and simple rule-based approach.

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model which is widely used. The approach leverages its application in understanding the context of the sentences and is used for various Natural Language Processing tasks. This approach gives the sentiment score ranging from -2 to +2 which is classified as positive, negative, and neutral. This method captures the complex meaning of the sentences and rates them accordingly.

TextBlob is a Python library which is built for executing certain NLP tasks and analyzing sentiments also. It is a similar approach to VADER, which calculates the polarity of the words using pre-trained lexicons and the polarity ranging from -1 to $+1$.

TextBlob method splits the news headlines into smaller phrases and the score is assigned based on its presence in the pre-defined lexicon.

The three approaches have their own applications, and which differs for each domain, the certainity of each sentimental analysers differs from each other as shown in the 5,

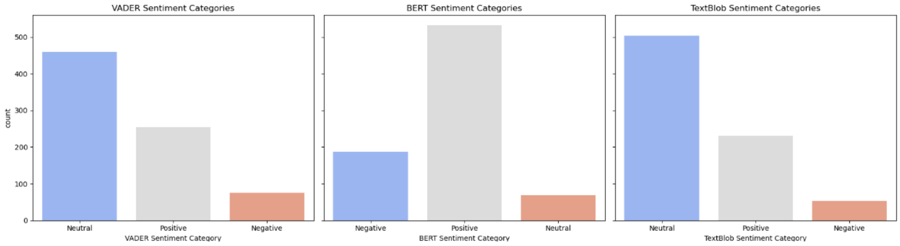

Figure 5: Visualizations of Sentimental Analyzers

(Juyal and Kundaliya, 2023)the importance of using hybrid-based approach is to bring out the parallel and direct correlation of the market data sentiments with the price changes of the cryptocurrency. Addressing the sensitivity, subtleness and context of the news headlines data and Reddit data, these three approaches were implemented. VADER and TextBlob approach are limited to only capture sentiments of the data which showcase high sentiment intensity in them, BERT model being able to capture complex contexts within the sentences might show probabilities in reacting the noises in the data. The news headlines and the Reddit data which is collected from the financial domain might not necessarily contain context related to financial domain, hence capturing the context of the sentence with non-financial nuanced meaning might also affect the sentimental scores. To maintain common range of the sentimental scores the values were scaled down using normalization and keeping the values ranging from 0 to 1, for which values closer to 0 represent negative sentiments, values closer to 1 represent positive sentiments, and the values that fall in middle ranges represent neutral. The average of the three sentimental scores are taken to capture the nuanced sentiment knowledge of the data, eliminating noise and which makes the data more reliable for the model,

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

Correlation analysis was conducted between the sentimental scores and the price changes to parallelly reflect the fluctuations of the price changes with the sentimental scores which makes the model capture the relationship between the sentiments and the price changes. Correlation scores of individual sentimental scores, averaged sentimental scores and hybrid sentimental scores is show below in table,

Sentiment Analysis	Price Changes Correlation
VADER	-0.037130
BERT	-0.108203
TextBlob	-0.026963
Average Sentiments	0.104131
Hybrid Sentiments	0.585954

Table 2: Correlation Analysis of the Sentimental Scores and Price Changes of Ethereum

Since volume factor of the Ethereum's financial data shows high correlation with the price changes in the market as show in the fig4, it was combined with the average sentimental scores to build a relationship between the sentimental data and the price fluctuations, and to also majorly reflect on the market's trading activity. By having this hybrid approach of creating a relationship between the text data and the financial data, the correlation coefficient was found to have a moderate positive relationship with the price changes which is evident for proving that hybrid approaches have more correlation with the price changes of Ethereum when compared to individual sentimental scores. This Hybrid based approach has its applications by addressing the limitations of each Sentimental Analyzer separately,

- Integrating rule-based, lexicon based (polarity) and context-based sentiments to have a holistic score of the market sentiment data.
- Enhances model capacity to draw relationships from the data, lowering its complexity factor.
- Inclusion of trading volume data in the sentiments adds real world context in the text data making the approach more effective for predicting the market movements.

fig 6 and fig 7 shows the average of the sentiments of the hybrid approach and the correlation analysis acts as an evaluation metrics for the hybrid approach implemented. (Aidoo and Ababio, 2023)To train the ML and DL models effectively with the time series data of Ethereum, stationarity analysis was conducted on the data which is to check the data's statistical properties change with respect to time, Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests were implemented to addressing the stochastic trends and deterministic trends. From the table 3 the negativity of ADF statistic, and the p value below 0.0002 is evident for rejection of the null hypothesis explaining the stationarity of the data also. And also the KPSS Statistic values is very much below all the critical values and with the p value of above 0.05 proves that the data do not need any stationarity conversion.

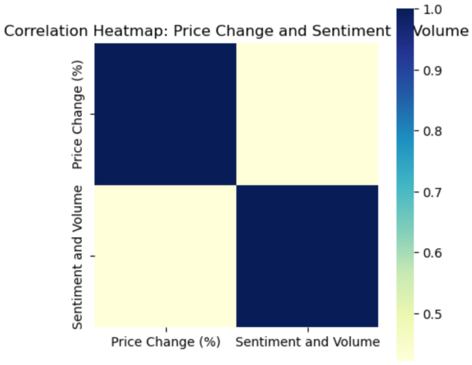

Figure 6: Correlation Analysis

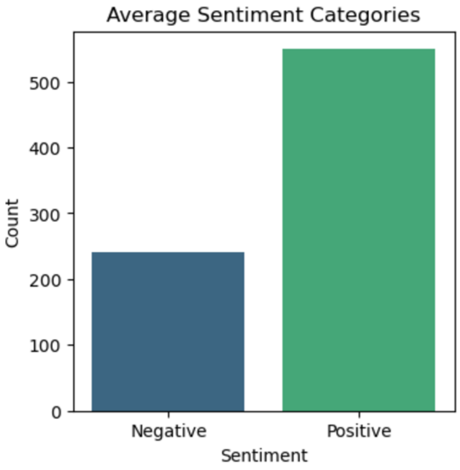

Figure 7: Average Sentiments

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

Table 3: Stationarity Tests

3.4 Modelling Approach and Evaluation Metrics

After enhancing the data quality and relationship factors with necessary feature engineering techniques, the ML and DL models are trained on this data to predict the price changes of the Ethereum. From the analysis 2 tree based models and LSTM models showed good scores in capturing the nonlinearity and complex relationship in the data. , Random Forest Regressor, XG Boost Regressor and LSTM models are chosen for doing the respective predictions by experimenting with different types of embeddings to represent the text data, TF-IDF, Word2Vec, BERT, and Sentence BERT embeddings are implemented on the text data and comparative analysis of their respective model's results are obtained and compared with each other. Random Forest Regressor and XG Boost regressor models help in easier interpretation of the predictions and helps in identifying the importances given to each feature which interprets the model's prediction more easily. And the trees are robust in handling the nonlinear relationships in the data, providing flexibility in capturing the feature interaction learning. LSTM has proven to provide good results on time series data, sequential data and in capturing long term relationships by analyzing the dependencies and is very useful in analyzing and understanding complex textual, sentimental and numeric features.

It is important to have an accurate model which is able to capture the variance in the dataset values, showing low error magnitude and providing an easy interpretable factor on the results obtained. (Munjal et al., 2024)The chosen evaluation metrics are Coefficient of Determination(R^2), Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and the time complexity of each embedding and for each for ML and DL models which assess the model's capacity in capturing the insights of the data, the model's performance is also assessed by visualizing the plot between actual and predicted values. The mentioned metrics evaluate the model capacity in the following ways;

1. R^2 – Evaluates the model's capacity in capturing the complex relationships between the data points and in learning the variance factor in the dataset.
2. RMSE – Evaluates the model's accuracy in the predictions made.
3. MSE – Model's performance is assessed by penalizing errors caused and highlighting the working capacity of the model.

The chosen text embeddings were chosen based on the nature of the data and assuming the context of the data to be very vague, TF-IDF which is simplest and an advantageous approach to represent each word's importance in the dataset, Word2Vec embeddings which represents the embeddings by capturing the semantic meaning in the short sentences like news headlines, BERT and Sentence BERT embeddings were implemented to have comparative analysis over the simplest to complex approaches and also to have embeddings which represents the linguistic issues and that captures the contextual information of the sentences. By implementing all these embeddings, the research study does a comparative analysis of the different embeddings on data which is nuanced in nature.

3.5 Explainable AI (XAI)

The primary aim of using XAI was to interpret the model's predictions, hence the features used were critically evaluated on their contribution in the model's predictions, particularly analyzing the features regarding the sentiments and the features which had the utmost impact on the Ethereum's prices. (Angelini et al., 2024) Analysis on the feature importances, partial dependence plots (PDP), Individual Conditional Expectation (ICE) and SHAP was implemented on the best working ML and DL model to explain the model's predictions. PDP and ICE was generated for hybrid sentiment features to analyze and generate the contribution and the impact of the hybrid sentimental features on the price predictions. As these insights would be effective in analyzing the effectiveness of creating a hybrid sentiment feature to predict the price changes. Additionally, the feature level contributions on each prediction and the interaction between each feature was implemented to understand on the factors which ae responsible for the volatile price changes and could act as an important baseline in the risk assessment factor in the cryptocurrency market. The utilized XAI methods and plots were helpful in analyzing the influence of the retrieved and feature-engineered features ensuring model's interpretability criteria.

4 Algorithm and Design Implementation

The research work addresses novelty by bringing in a unique and hybrid approach in using the sentiments from the discussion forums of Reddit and global news headlines also as an important additional factor in predicting the Ethereum's price changes. The outline of the algorithm working is shown below:

Algorithm 1 Hybrid Sentiment Analysis

```
Input: Ethereum financial data, news headlines
Output: Positive correlation between sentiments and price changes
for each row in dataset do
    Calculate sentimental scores using VADER, BERT and TextBlob
    Compute the average of the sentimental scores
    Analyse the correlation scores between individual sentiments and price changes
end for
Hybrid Sentimental Scores: combining average sentimental scores and volume feature
    together
Comparison between hybrid sentimental scores and individual sentimental scores
end =0
```

The research was conducted with a system configuration of 8GB RAM, with all the computations performed on the CPU in a Windows operating system. Python 3.12.4 was the programming language used in the development of the research work in Jupyter Notebook environment created from Anaconda Navigator (anaconda3). Some of the important Python libraries used for model implementation and sentimental analysis are VADER, TextBlob, Hugging Face Transformers, Scikit-learn, TensorFlow. Pandas, NumPy, Sklearn, Matplotlib and Seaborn for data processing, transformation and visualization.

SHAP for XAI and APIs used in collection of the data are Reddit, Media retrieval, GDELT and Kraken – APIs.

5 Implementation

The research work aims in bringing the best ML or DL model which is better at capturing the complex relationship between the hybrid sentiments and the financial data parameters in predicting the overall price changes of the Ethereum cryptocurrency. The dataset after necessary cleaning, transformation, feature engineering, and exploratory data analysis, was ready for the model training. The text data was converted into various embeddings to have a comparative analysis of the text data based on their nature. By this comparison, the research work was able to arrive at results that the text data from these sources like Reddit discussion forums and news headlines. There were four types of embeddings which were chosen for representing the text data to address the nature of the data in different forms.

1. TF-IDF embeddings which is a simple and yet a very useful technique in addressing the importance of the words used in the data, the vectorizer uses maximum of top 5000 words based on its importance and provides a high dimensional sparse representation of the text in the for vectors.

2. Word2Vec embedding was implemented to address the semantic similarity of the words used in these platforms by taking a vector size of 100, minimum count of each word appearing was set to 1 and a window size of 5 for each word and its corresponding neighboring words. This method was useful in capturing the semantic meaning of the words represented in the form of dense vectors.

3. BERT embedding was implemented to capture the overall context of the sentences using 'yiyanghkust/finbert-tone' model which is fine tuned for sentimental analysis of financial data.

4. SBERT embedding - 'paraphrase-MiniLM-L6-v2 'represents the context of the sentences on a high level with enhanced quality which is mainly to obtain optimized format of sentence representation.

BERT and SBERT embedding are mainly implemented for their context aware in the sentences and for capturing the relationship between different sentences used in the data. And both the models are implemented using transfer learning, which was pretrained on massive datasets, and allowing the model to use the gained knowledge to better understand the given data and leveraging fine tuning process.
Initially the tree-based models were selected based on the analysis, Random Forest regressor and XG Boost regressor models were implemented. The models were trained by taking each embedding separately and were critically analyzed for the nature of the data involved. The dataset was split into 80 percent training and 20 percent for testing for both the models with reproducible factor kept constant as 42. The number of estimators used for Random Forest regressor model was 100 and the results were obtained accordingly.

And Grid Search was used as hyperparameter optimizing factor while training the XG Boost model and was initialized as follows, Trees to be used - (100, 150, 200), maximum depth - (3,5,7), learning rate scheduler was (0.01, 0.1, 0.2) and fraction of samples and features used were (0.8,1.0) and (0.3, 0.5, 0.8) and the grid search used 3-fold cross-validation by returning the best model with the right and optimized hyperparameters. Grid search was helpful in finding the optimized parameters, with no guesswork involved and always serves as a reliable approach in tuning hyperparameters of the model.

Likewise, a DL model – LSTM was also trained with the same data split ratio as the ML models and had to undergo the same process of training on the individual embeddings with all the other parameters. And the model was built with 50 input nodes, 0.2 of dropout layer was used as regularization factor considering the size of the dataset, with ReLU activation layer and with one output neuron which gives the predictions. Adam optimizer was implemented for having an accurate and adaptive learning rate according to the data.
The model produced its best results when it was trained for 10 epochs and a batch size of 8. The three models were critically compared against each other to produce the best results and the XAI was implemented for the model that produced the best results.

6 Evaluation

The critical comparative analysis of the ML and DL models are evaluated based on the regression evaluation metrics: R^2, RMSE, MSE, visualization of the model's actual vs predicted graph to know the working capacity of the model in capturing the variance and complex relationships and their respective time complexity involved in training.

6.1 Case Study 1-Results

Results of Random Forest Regressor model trained on different embeddings to predict the price changes

Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
TFIDF Embeddings	**0.000395**	**0.977451**	**0.019866**	**2.814366**
Word2Vec Embeddings	0.001071	0.938789	0.032731	6.700570
BERT Embeddings	0.002070	0.881754	0.045493	53.109120
SBERT Embeddings	0.001881	0.892502	0.043376	26.448556

Plot of model's actual vs predicted was plotted to check the accurateness of model's predicted values with the actual values

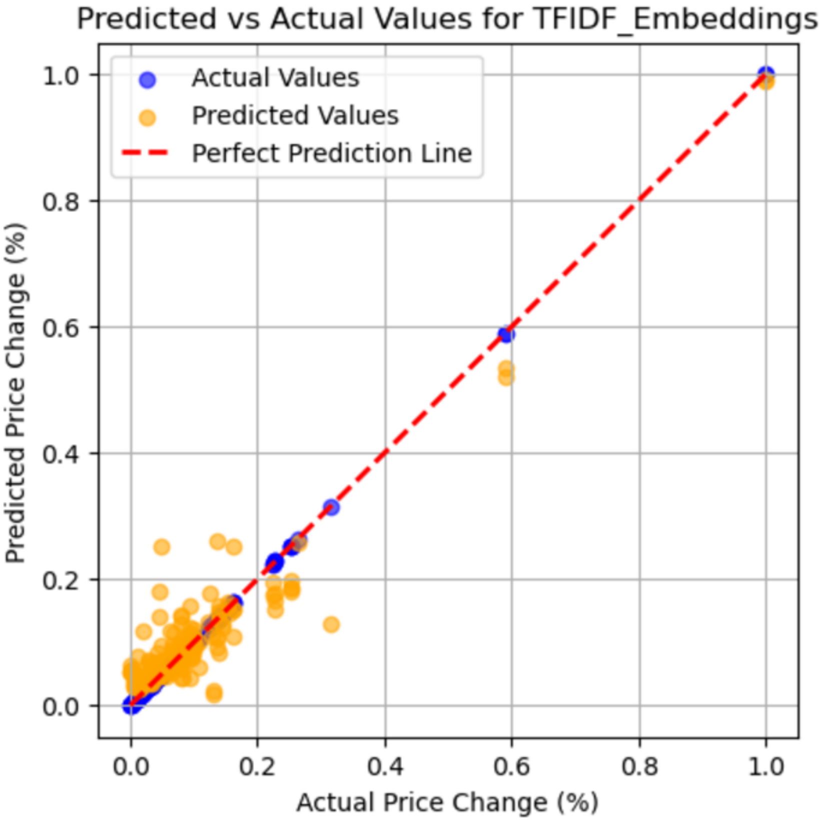

Figure 8: RF Regressor Results for TF-IDF

6.2 Case Study 2-Results

Evaluation results of XG Boost Regressor model trained on different embeddings to predict the price changes

Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
TFIDF Embeddings	**0.000313**	**0.982115**	**0.017693**	**1052.430193**
Word2Vec Embeddings	0.000784	0.955207	0.028	294.155899
BERT Embeddings	0.001922	0.890197	0.043839	1857.387344
SBERT Embeddings	0.001581	0.909643	0.039768	992.234312

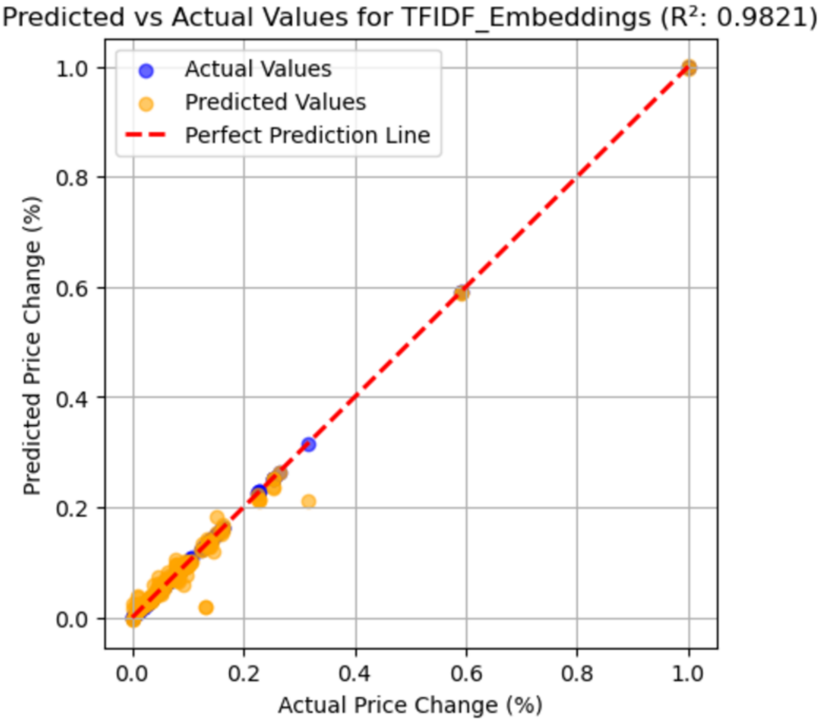

Figure 9: XGB Regressor Results for TF-IDF

6.3 Case Study 3-Results

Evaluation results of LSTM model trained on different embeddings to predict the price changes

Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
TFIDF Embeddings	0.008227	0.529958	0.090702	1809.632197
Word2Vec Embeddings	**0.007221**	**0.587432**	**0.084976**	**1815.184139**
BERT Embeddings	0.017332	0.009722	0.131653	1822.326937
SBERT Embeddings	0.016412	0.062288	0.128111	1828.319779

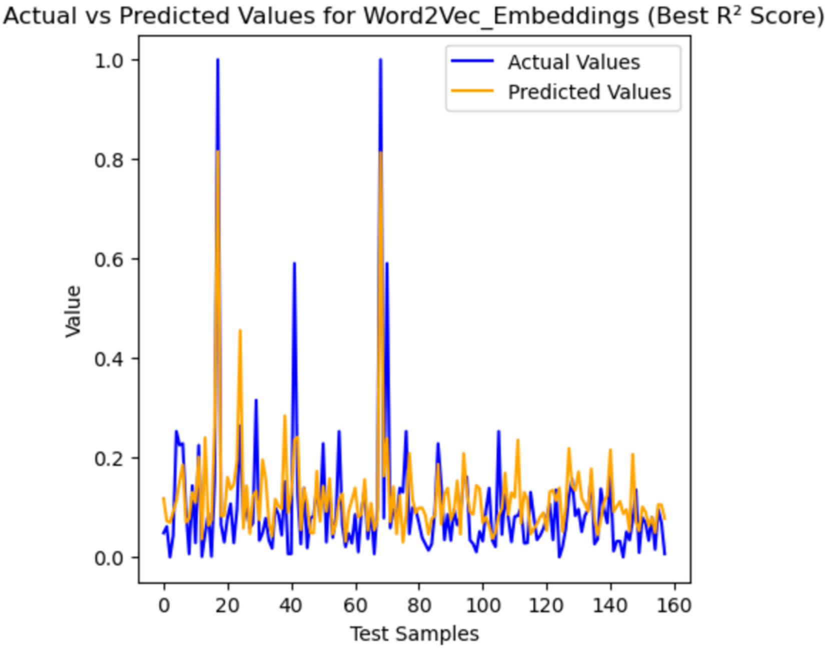

Figure 10: LSTM Results for Word2Vec

6.4 Case Study 4-Results

Evaluation results of XG Boost Regressor model trained on only the hybrid sentiment without any other financial features to predict the price changes and plot of the same is plotted below:

Embeddings	Mean Squared Error	R^2 Score	Root Mean Squared Error	Time Taken (seconds)
TFIDF Embeddings	0.009182	0.475386	0.095823	0.61

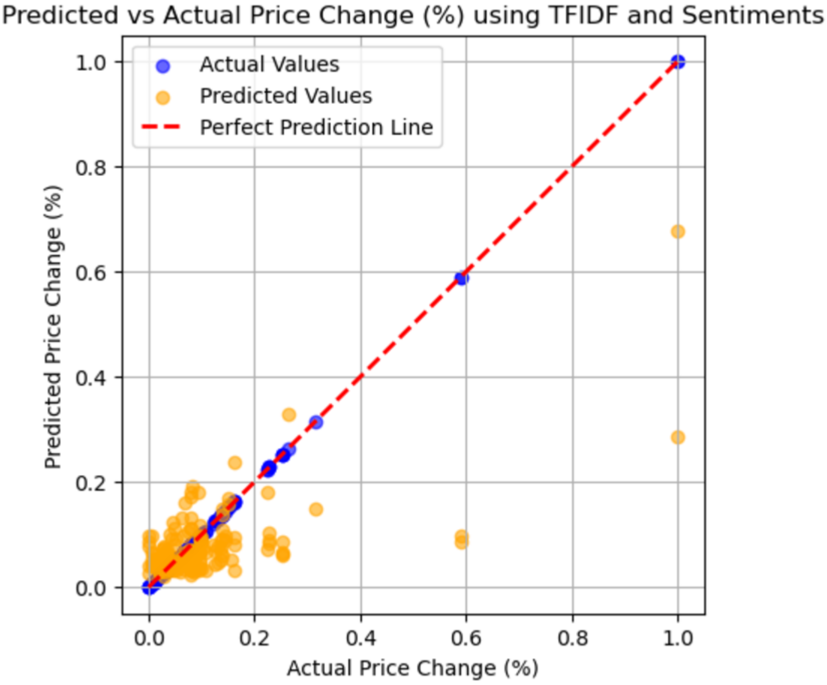

Figure 11: XGB model Results with only Average Sentiments and Hybrid Sentiments

6.5 XAI Results on XG Boost Regressor

To analyze the strength of the hybrid sentiment features in the model training, the complexity and interaction of the hybrid sentiment feature with other features was evaluated using explainable AI, PDP and ICE plots. PDP and ICE plots helped to visualize the contribution of the average sentiments and hybrid sentiments in predicting the price changes of Ethereum and was good enough to reveal the maximum contribution of the individual sentiments, average sentiments and hybrid sentiments revealing their strengths in the learning process of the model. The below visualizations show the individual contributions of the average sentiments and hybrid sentiment features in the model training.

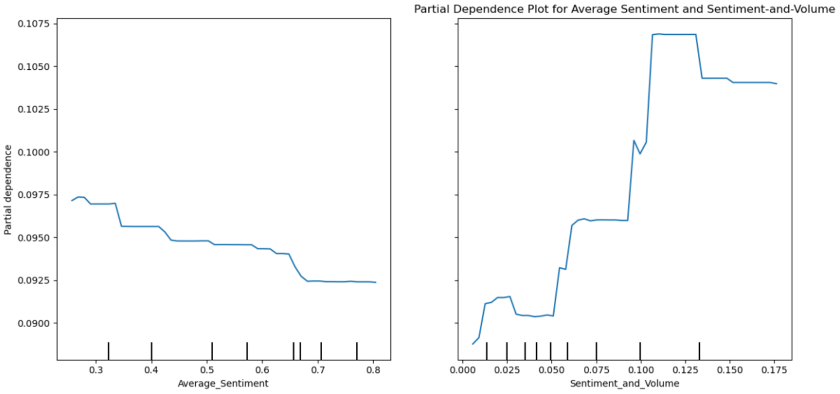

Figure 12: PDP visualization of the Average and Hybrid sentiments

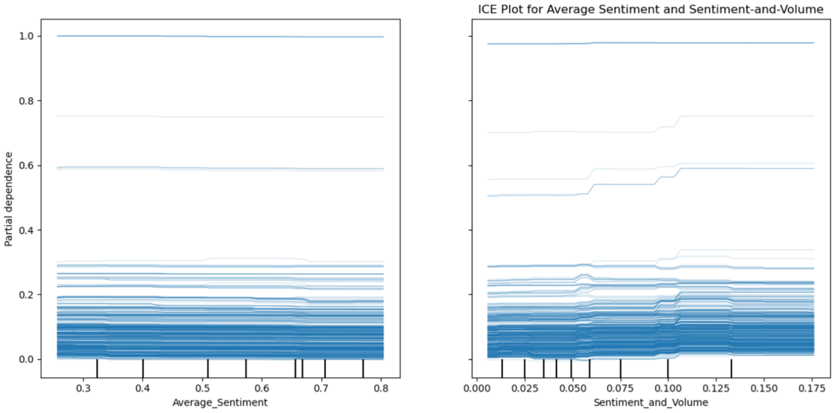

Figure 13: ICE visualization for Average and Hybrid Sentiment

The research work also evaluated the impact of the hybrid sentiment feature on the first instances 25 percent of the test dataset by leveraging Shap Which was helpful in determining the effectiveness of the feature in the model training, and the results were

"Positive SHAP contributions: **39 out of 40**"

Shap provided insights on each feature within the dataset interacted with each other, how each feature contributes to the model's prediction for each instance in the dataset. The non-linearity in the data was explained using Shap by identifying the strengths of each feature to itself and others and helps in demonstrating the nonlinearity exhibited by the dataset during the model training and prediction. The insights provided by the Shap analysis purely deals with only the features as that remains the major contribution of the research work.

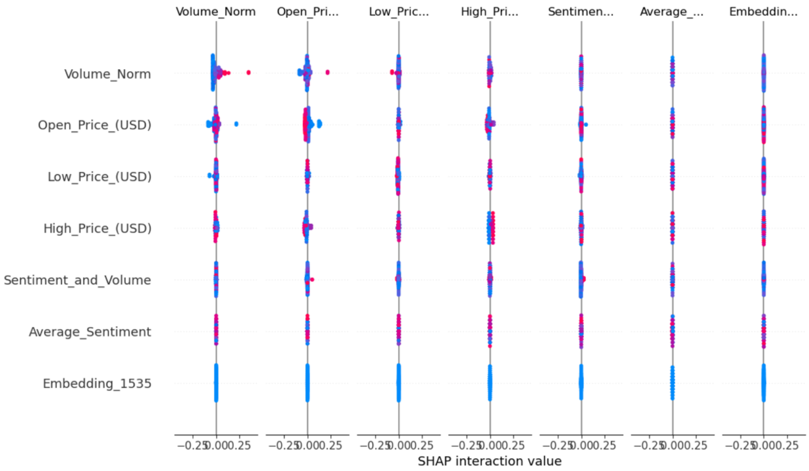

Figure 14: SHAP interaction value

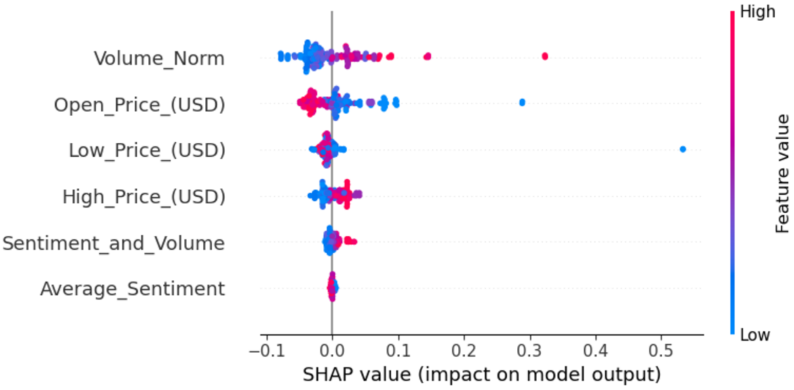

Figure 15: Impact of model's predictions using SHAP values

These evaluations were implemented on the model to explain the interaction of the features in the model training and also explains the transparency of each feature to the model's predictions.

6.6 Discussion

6.6.1 Model Results Interpretation

The comparative analysis between the ML and DL models led to conclusions regarding the performance of model depending on the nature of the model. Random Forest model provided good results in capturing the relationship between the financial data and the sentimental data in predicting the price changes, particularly the model performed well on the dataset which TF-IDF embeddings, which conveys that a tree models like Random Forest shows its strength in capturing the statistical strength of the words rather than other embeddings which uses context of the sentence. XG Boost model outperforms the other two models in the analysis by learning the variance involved in the dataset very well and the model also works exceptionally good for the dataset which uses TF-IDF embedding. XG Boost model, which is also a tree-based boosting model shows its strength in capturing the short and seasonal trends shown in the dataset. Similar to Random Forest, XG Boost model also shows good results with embeddings which show statistical relationship rather than contextual relationships. LSTM model yields average results when compared to other two models and the model also shows good results with statistical embeddings like TF-IDF and Word2Vec, the reason for the average performance of the model is that the model shows its strength in capturing only the long term dependencies or sequences and fails to perform best on seasonal or short trends. The model shows good results on dataset that includes Word2Vec embeddings and yields negative results on contextual embeddings like BERT and SBERT. And the final case study is evident that the hybrid sentiment data along with the financial parameter shows good results when compared to only the average sentiment data.

6.6.2 XAI Results Interpretation

The PDP and ICE plotted for average sentiment and hybrid sentiments visualize that the hybrid sentiments show a high nonlinear increase in the trend of capturing the variance of the price changes compared to the average sentiments feature. The increase in the trends of the hybrid sentiment compared to the sentiments shows that the model effectively gives more importance to hybrid sentiment more than average sentiments.

The Shap feature interaction plot and summary plot visualizes that the financial parameters contribute more to the model's predictions more than the sentiment data, particularly volume of the trading investors shows high interactions in the model's predictions. However, the research work focuses on leveraging the importance of the hybrid sentiment features with comparative analysis on the average and individual sentimental scores. Hence the research work shows a strong analysis that hybrid sentiments contribute most to model's prediction and help in better learning and training of the dataset.

6.6.3 Comprehensive Analysis of the Results

The research study arrives at the results conveying that the hybrid features help in reducing noise in the text data, arriving at a holistic approach on the sentiments. And with the consideration of sentiments in the model training was helpful determining the strength of machine learning and deep learning in the financial market which is very useful in assessing the risk factors of cryptocurrency. With this the research, the study was able to accomplish the objectives, which was to assess the market investors psychology by building a strong relationship in the sentiments of the data with the price changes. Arriving at high quality sentiments from the text data which is obtained from a subtle source is very important, hence addressing the importance of deploying sentiment on the financial data and assessing the quality of the sentiment data is ample for the research work to conclude its results to obtain accurate and precise Ethereum's price predictions.

7 Conclusion and Future Work

Considering the importances of cryptocurrencies and its market in today's world, Ethereum was chosen for the research work due to various financial and global factors. Research work aimed solving some of the business perspectives involved in cryptocurrency trading by analyzing the challenges that are experienced by the investors. Hence the problem was narrowed and involved more complications to analyze the sentiments from the news headlines and online Reddit discussion forums. To train ML and DL models from the retrieved sentiments and financial data was the primary aim of the research work. Dealing with the complexities in analyzing the sentiments retrieved from such subtle platforms leads to creation of new and separate hybrid sentimental features and making these features align with other financial parameter and train the models effectively by capturing the utmost insights, patterns, and variance from the dataset was another important area the study focused on. Arriving at critical analyzing of the results and assessing the importance of the created and existed features helps the research work to succeed in the focused goals and gives room for further technical and business discussions about the work.

XG Boost model performed well in utilizing the dataset features, capturing the non-linear relationships from the dataset and outperformed the other two models. Necessary visualizations and explanations were implemented on the model to explain the working of the created features and other features, which evaluates the created hybrid sentimental feature and also serves as an extra layer in explaining the working of the model or transparency of the model in arriving at the accurate predictions. The PDP and ICE plots results provided the interaction of the hybrid sentimental and average sentiment features separately in the price predictions and was also helpful in a comparative analysis. The Shap values provided the model's importance on each feature and the feature interaction for the model predictions, this stage explained that the created had a more importance than the average and individual sentimental features. The study explains the importance of using high quality sentiments in predicting the prices of Ethereum, as the sentiments of the news headlines or discussion forums will be the major contributor in impacting the decisions made by the investors. As the visualizations show that volume of the investors on a daily basis have high influence in the price changes and in creating more volatility in the market.

The research work addresses the seasonal trends, short events and hence chooses Ethereum for the volatility it has shown over the past few years. As the results were promising to show that the high-quality sentiments were good enough to predict the prices involved in a short real time span, experimenting with long time dependencies and sequences could lead to more training of more complex deep learning models and leads to better and complex analysis. Additionally, the dataset used in the study was not identified for any biases as the dataset used was assumed to be utilized only for price predictions where bases cannot influence the parameters but in the real time the prices show biases for sentiments which come from top investors. With all these additional factors the research work can be further improved utilizing the strengths of the datasets discovered leading to more robust risk assessment tool in the market.

References

Aidoo, D. and Ababio, K. A. (2023). Modeling bitcoin prices and returns using arima model, *International Journal of Innovation and Development* **Special Edition (December 2023)**. Corresponding author's email: aidoodavid86@gmail.com.

Akhand, M. N. T., Habib, M. A. and Alam, K. M. R. (2023). Analyzing cryptocurrency price trends for real-time price predictions, *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, IEEE, Cox's Bazar, Bangladesh.

Angelini, M., Blasilli, G., Lenti, S. and Santucci, G. (2024). A visual analytics conceptual framework for explorable and steerable partial dependence analysis, *IEEE Transactions on Visualization and Computer Graphics* **30**(8).

Armin, A., Shiri, A. and Bahrak, B. (2022). Comparison of machine learning methods for cryptocurrency price prediction, *2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, IEEE, IEEE, Mazandaran, Iran.

Aslam, N., Rustam, F., Lee, E., Ashraf, I. and Bernard, P. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model, *IEEE Access* **10**. Received March 19, 2022, Accepted April 1, 2022, Published April 7, 2022, Current Version April 18, 2022.

Aslim, M. F., Firmansyah, G., Akbar, H., Tjahyono, B. and Widodo, A. M. (2023). Utilization of lstm (long short-term memory) based sentiment analysis for stock price prediction, *Asian Journal of Social and Humanities* **1**(12): 1241–1254.

Bhadula, S., Kartik, D. and Gupta, D. (2024). An explainable ai regression model for gold price prediction, *Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Amrita School of Computing, Amrita Vishwa Vidyapeetham, IEEE, Bengaluru, India.

Cruz, L. F. S. A. and Silva, D. F. (2021). Financial time series forecasting enriched with textual information, *Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, São Carlos, Brazil, pp. 480–485.
URL: *https://doi.org/10.1109/ICMLA52953.2021.00066*

Farimani, S. A., Jahan, M. V. and Fard, A. M. (2022). From text representation to financial market prediction: A literature review, *Information* **13**(10): 466.
URL: *https://www.mdpi.com/1999-5903/13/10/466*

Farimani, S. A., Jahan, M. V. and Fard, A. M. (2024). An adaptive multimodal learning model for financial market price prediction, *IEEE Access* **12**: 121846–121859.

Fior, J., Cagliero, L. and Garza, P. (2022). Leveraging explainable ai to support cryptocurrency investors, *Future Internet* **14**(9): 251.

Gontyala, S. P. (2021). *Prediction of cryptocurrency price based on sentiment analysis and machine learning approach*, Master's thesis, National College of Ireland. MSc Research Project.

Goodell, J. W., Ben Jabeur, S., Saadaoui, F. and Nasir, M. A. (2023). Explainable artificial intelligence modeling to forecast bitcoin prices, *International Review of Financial Analysis* **88**: 102702.

Gupta, A., Jain, H., Zope, B., Vyas, D., Mishra, S., Buchade, A., Nale, P. and Bidwe, R. V. (2023). Cryptocurrency prediction and analysis between supervised and unsupervised learning with xai, *2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, IEEE, pp. 1–6.

Jaiswal, S., Srivastava, S., Garg, S. and Singh, P. (2023). Effect of news headlines on gold price prediction using nlp and deep learning, *2023 International Conference on Artificial Intelligence and Applications (ICAIA)*, IEEE, pp. 1–9.

Juyal, P. and Kundaliya, A. (2023). A comparative study of hybrid deep sentimental analysis learning techniques with cnn and svm, *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, IEEE, pp. 596–600.
URL: *https://ieeexplore.ieee.org/document/10263883*

Munjal, G., Khandelwal, V. and Varshney, H. (2024). Sentiment analysis based stock price prediction using machine learning, *Proceedings of the 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Amity University, IEEE, Noida, Uttar Pradesh, India.

Oikonomopoulos, S., Tzafilkou, K., Karapiperis, D. and Verykios, V. (2022). Cryptocurrency price prediction using social media sentiment analysis, *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, IEEE, Thessaloniki, Greece.

Passalis, N., Avramelou, L., Seficha, S., Tsantekidis, A., Doropoulos, S., Makris, G. and Tefas, A. (2022). Multisource financial sentiment analysis for detecting bitcoin price change indications using deep learning, *Neural Computing and Applications* **34**: 19441–19452. Received: 2 December 2021; Accepted: 1 June 2022; Published online: 3 July 2022.
URL: *https://doi.org/10.1007/s00521-022-07509-6*

Raheman, A., Kolonin, A., Fridkins, I., Ansari, I. and Vishwas, M. (2024). Social media sentiment analysis for cryptocurrency market prediction, *Proceedings of the Autonio Foundation Conference*, Autonio Foundation.

Rateb, M. N., Alansary, S., Elzouka, M. K. and Galal, M. (2024). Cryptocurrency price forecasting implementing sentiment analysis during the russian-ukrainian war, *Preprint* . Available under a Creative Commons Attribution 4.0 International License.
URL: *https://doi.org/10.21203/rs.3.rs-3835106/v1*

Sahal, R. (2022). *Predicting optimal cryptocurrency using social media sentimental analysis*, Msc research project, National College of Ireland. Supervisor: Dr. Catherine Mulwa.

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): 791–797.

Singh, M., Juneja, A., Jakhar, A. K. and Pandey, S. (2023). Machine learning based framework for cryptocurrency price prediction, *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, IEEE, IEEE, Una, India.

Yu, D. (2022). Cryptocurrency price prediction based on long term and short term integrated learning, *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, IEEE, IEEE, Ningbo, Zhejiang, China.

Zubair, M., Aurangzeb, K., Ali, J., Alhussein, M., Hassan, S. and Umair, M. (2024). An improved machine learning-driven framework for cryptocurrencies price prediction with sentimental cautioning, *IEEE Access* **12**. Received 6 January 2024, Accepted 24 January 2024, Published 19 February 2024.
URL: *https://doi.org/10.1109/ACCESS.2024.3367129*