International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Bitcoin Price Prediction using Twitter Sentiment Analysis

[1]Himanshu Bute, [2]Abhyuday Singh, [3]Shlok Nandurbarkar, [4]Shivali Amit Wagle, [5]Preksha Pareek

Submitted: 20/12/2023 **Revised**: 31/01/2024 **Accepted**: 10/02/2024

Abstract— This work explores predicting Bitcoin prices using sentiment analysis on Twitter. Leveraging machine learning and rule-based methods, the study correlates social media sentiment, especially on Twitter, with dynamic Bitcoin prices. The dataset combines historical Bitcoin prices from Yahoo Finance and relevant Twitter data. Preprocessing involves cleaning tweets, calculating sentiment scores, and merging datasets. Results indicate that SGD regression and ridge regression achieve the best performance with a Validation-MAPE of 8.45%. Despite Bitcoin's volatility, the study highlights the potential of sentiment analysis in forecasting values, shedding light on the intricate relationship between social media sentiments and cryptocurrency markets.

Keywords*—Bitcoin, Sentiment analysis, preprocessing*

Introduction

The frequency of social media posts, news, and public opinions is found to be directly connected to the price movement of current digital assets like cryptocurrencies. The main factor behind the price movement has been found in Tweets on Twitter (now X) and Reddit. Bitcoin is one of the first cryptocurrencies that has shown significant growth in market dominance in recent years [1]. That being said, it has also been found that the variation in the prices of BitCoin has been very dynamic and has been heavily influenced by the opinions of powerful people, public sentiments on social media, and political scenarios. The Cryptocurrencies are directly related to blockchain frameworks and device its various properties like transparency and decentralization [3].

In this study, we develop an end-to-end model for predicting the price movement of these Cryptocurrencies based on sentiment analysis from social media. Researchers have shown significant interest in examining public sentiment through platforms like Twitter and Reddit. [1][19]. Additional elements influencing cryptocurrency prices may include transaction costs, mining complexities, alternative coin options, etc. [3]. After the success of Bitcoin, numerous alternative cryptocurrencies, commonly referred to as 'altcoins,' have emerged. [4]. Investors are increasingly utilizing cryptocurrencies as short-term assets, emphasizing the significance of accurate predictions. [13]. The attractiveness of this market lies in the fact that the technology employed in cryptocurrency mining offers a viable alternative to conventional markets like gold. [5]. However, Cryptocurrency investment is not visible due to the market’s inconsistent aspect and volatility of high prices [3].

Using sentiment analysis to forecast bitcoin values needs two separate procedures. Fundamentally, there are two ways to accomplish sentiment analysis [20]: using machine learning or rule-based approaches. The rule-based method categorizes tweets using a preset set of rules and a vocabulary of words. This is exemplified by the rule- and lexicon-based sentiment analysis tool VADER (Valence Aware Dictionary and Sentiment Reasoner), which requires no training. Positive, negative, and neutral sentiment scores are generated; these ratings are then normalized between 0 and 1. On the other hand, tweets are analyzed using machine learning techniques that use algorithms like Naive Bayes, logistic regression, and linear support vector machines. The best method for accurately predicting price changes for Bitcoin was determined

Symbiosis Institute of Technology, (Pune campus), symbiosis International Deemed University, Pune 411215, India

himanshu.bute.mtech2023@sitpune.edu.in,

abhyuday.singh.mtech2023@sitpune.edu.in,

shlok.nandurbarkar.mtech2023@sitpune.edu.in,

Shivali.wagle@sitpune.edu.in,

Preksha.pareek@sitpune.edu.in

Corresponding author: Shivali Amit Wagle (shivali.wagle@sitpune.edu.in)

to be logistic regression. The CEPM model performs better than the individual models, according to experimental analysis of Twitter datasets gathered during the COVID-19 era [4]. However, the task's difficulty might be related to the numerous variables and unpredictabilities that interplay in the markets, such as prevailing social, political, and economic trends. It is very challenging, but not impossible, to predict market price changes consistently [5].

A. Related Work

First, In [1], the researchers utilized an RNN-LSTM model to forecast changes in the price of bitcoin using previous price data and data on user attitude. In predicting the direction of Bitcoin price fluctuations over the next 24 hours, the model had an accuracy of 85%. This shows that using public mood analysis to forecast Bitcoin values may be useful.

The MM-LSTM model did well, scoring an impressive 95% accuracy in predicting Bitcoin prices a day ahead, as mentioned in [3]. This is a big jump compared to other deep learning models, usually hitting around 80-85% accuracy. The research suggests that MM-LSTM could be a promising new way to predict Bitcoin prices, capturing both short-term and long-term trends in cryptocurrency pricing across various cryptocurrencies.

Meanwhile, in [6], researchers found that their model was about 64.2% accurate in predicting the daily trend in Bitcoin prices. They also discovered that sentiments expressed on Twitter can be a handy indicator of Bitcoin price changes – positive sentiments correlating with price increases and negative sentiments with drops.

The FGM(1,1) is a potential new technique for forecasting the price of blockchain coins, the authors write in [11]. However, they also point out that the FGM(1,1) is still being developed and that additional studies are required to confirm its efficacy across a bigger dataset and a longer time frame.

According to [14], the model's mean absolute percentage error (MAPE) for Bitcoin, Ethereum, and Litecoin was 2.3%, 2.7%, and 3.1%, respectively. This contrasts well with the MAPE calculated using a conventional ARIMA model, which is 3.1% for Bitcoin, 3.5% for Ethereum, and 4.0% for Litecoin. The study offers proof that deep learning and neural networks can be used to anticipate Bitcoin prices reasonably well.

A. Twitter Data

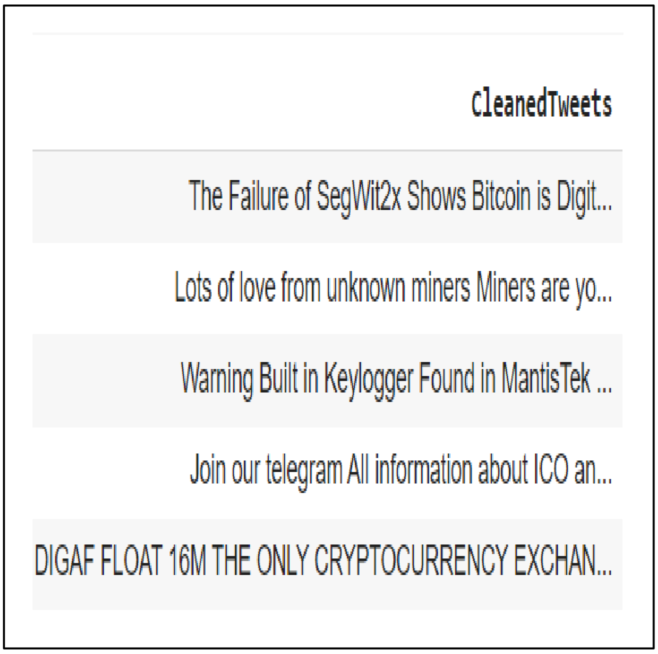
The *tweet-preprocessor* library available in Python was used for pre-processing tweets. With the help of this library, the non-conventional data types related to Twitter like URLs, Emoji, and Mention were converted to their equivalent text counterparts. Further, any unconventional characters were cleared out of the tweets by matching the tweets with the following *regular expression* and then replacing them with a blank space.

```
Regex Pattern = (@[A-Za-z0-9]+)|([^0-9A-
Za- z \t])|(\w+:\/\/\S+)
```

After doing all the processes mentioned previously a new column was prepared “Cleaned Tweets” (as per Fig. 1) which contained the processed tweets.

Dataset Description

The dataset in this study consists of two data sources. One data source is the price data of Bitcoin, acquired from Yahoo Finance. This data originally contained the “date”, “time”, “open-price”, “high-price”, “low-price” and “close-price”. However, we have modified this data to only include “date- time” and “closed price” as these were the only columns we required in our study.



CleanedTweets

The Failure of SegWit2x Shows Bitcoin is Digit...

Lots of love from unknown miners Miners are yo...

Warning Built in Keylogger Found in MantisTek ...

Join our telegram All information about ICO an...

DIGAF FLOAT 16M THE ONLY CRYPTOCURRENCY EXCHAN...

Fig. 1. CleanedTweets Column

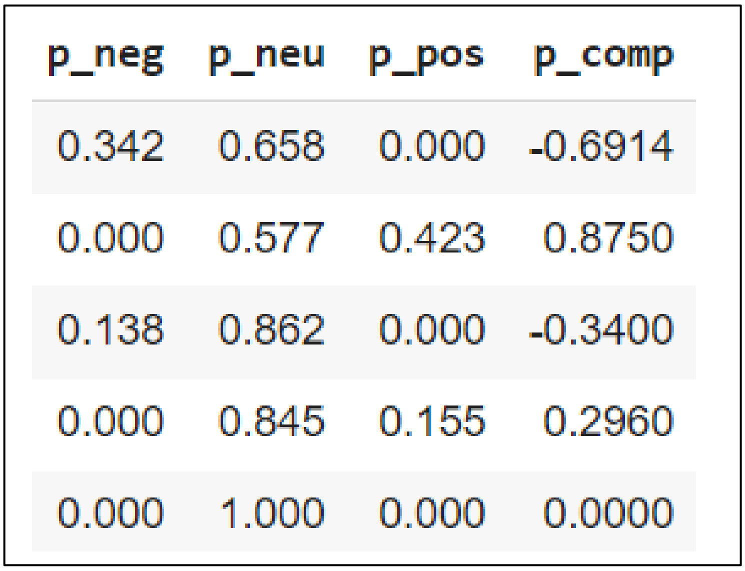
The second part of the dataset consists of the scrapped Twitter data for the date range similar to that of the Bitcoin dataset. This Twitter dataset therefore consists of two columns "Date-Time" and "Tweets". The tweets column only contains scrapped tweets that matched the query of keywords similar to Bitcoin, crypto, etc, and tweets from the accounts that have more than 100K followers, i.e. the tweets that actually may have an impact. The description of the dataset is given in Table 1.

TABLE I. Dataset Description

Table Name	**Columns**
Bitcoin Dataset	Date-Time
	Open Price
	High Price
	Low Price
	Close Price
Twitter Dataset	Date-Time
	Raw Tweets (Text)

Data Pre-processing

The Twitter Dataset in totality had **25,64,350** rows which was divided among three .csv files having around **8,54,783** rows each. Thus all three files were concatenated together to be used as a single Dataframe. The VaderSentiment packages available in Python were used to calculate sentiment scores for the tweets. The scores acquired were 'p_neg', 'p_neu', 'p_pos', and 'p_comp' which stand for the negative, neutral, positive, and compound as shown in Fig. 2.



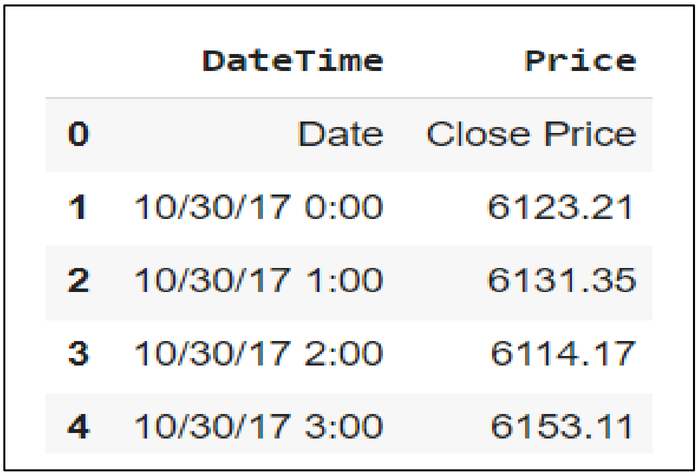
p_neg	p_neu	p_pos	p_comp
0.342	0.658	0.000	-0.6914
0.000	0.577	0.423	0.8750
0.138	0.862	0.000	-0.3400
0.000	0.845	0.155	0.2960
0.000	1.000	0.000	0.0000

Fig. 2. Sentiment Scores Example.

B. *Bitcoin Data*

The Bitcoin Dataset originally contained 5 columns, out of which 3 columns namely “open-price”, “high-price”, and “low- price” were removed as they served no purpose in the final predictions in our case, which is represented in Fig. 3

A. *Density Plots*



	DateTime	Price
0	Date	Close Price
1	10/30/17 0:00	6123.21
2	10/30/17 1:00	6131.35
3	10/30/17 2:00	6114.17
4	10/30/17 3:00	6153.11

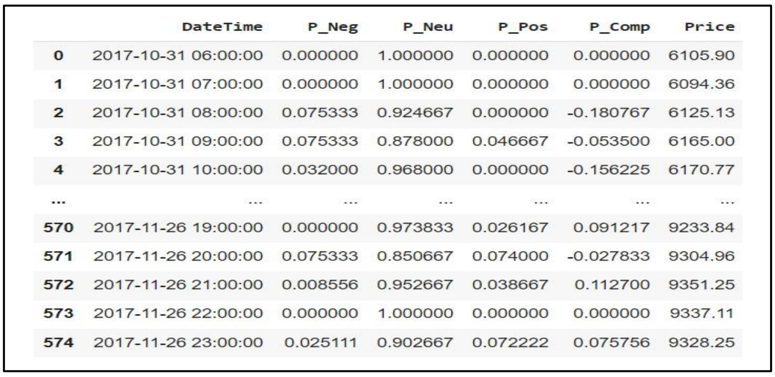
Fig. 3. Bitcoin Dataset

C. *Merging of Datasets*

After processing both the Twitter and bitcoin datasets the rows were grouped based on date-time columns in both the datasets and then joined together over a date-time column (as shown in Fig. 4).

Finally, both the tweet columns were dropped out leaving us with the columns mentioned below:

- Date-time
- P_neg
- P_neu
- P_pos
- P_comp
- Price



	DateTime	P_Neg	P_Neu	P_Pos	P_Comp	Price
0	2017-10-31 06:00:00	0.000000	1.000000	0.000000	0.000000	6105.90
1	2017-10-31 07:00:00	0.000000	1.000000	0.000000	0.000000	6094.36
2	2017-10-31 08:00:00	0.075333	0.924667	0.000000	-0.180767	6125.13
3	2017-10-31 09:00:00	0.075333	0.878000	0.046667	-0.053500	6165.00
4	2017-10-31 10:00:00	0.032000	0.968000	0.000000	-0.156225	6170.77
...	...	...	...	...	...	...
570	2017-11-26 19:00:00	0.000000	0.973833	0.026167	0.091217	9233.84
571	2017-11-26 20:00:00	0.075333	0.850667	0.074000	-0.027833	9304.96
572	2017-11-26 21:00:00	0.008556	0.952667	0.038667	0.112700	9351.25
573	2017-11-26 22:00:00	0.000000	1.000000	0.000000	0.000000	9337.11
574	2017-11-26 23:00:00	0.025111	0.902667	0.072222	0.075756	9328.25

Fig. 4. Merged Dataset

Exploratory Data Analysis

We performed various analyses of the data to analyze the trends in the price data concerning other features. We also analyzed the density of all the scoring parameters using a Density plot. All the various plots are shown in Fig. 5,6, 7, 8, and 9

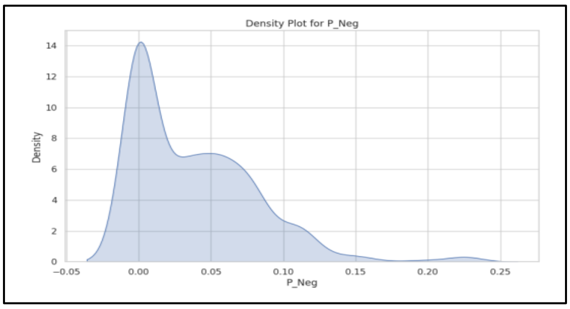


Fig. 5. Density Plot for P_Neg

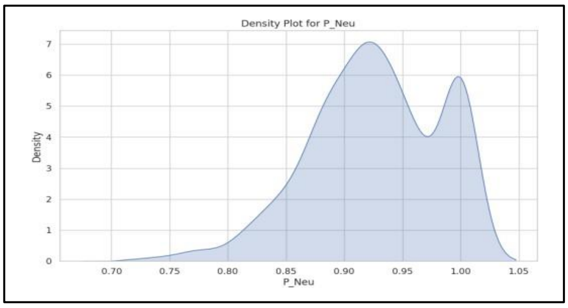


Fig. 6. Density Plot for P_Neu

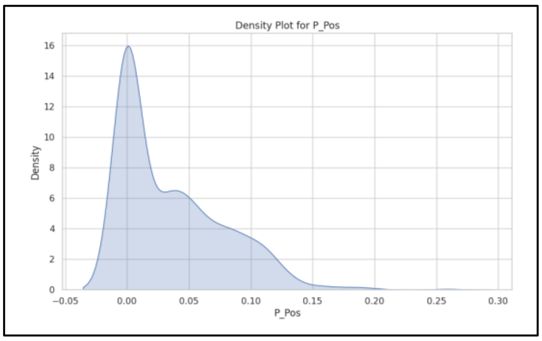


Fig. 7. Density Plot for P_Pos

B. *Line Chart*

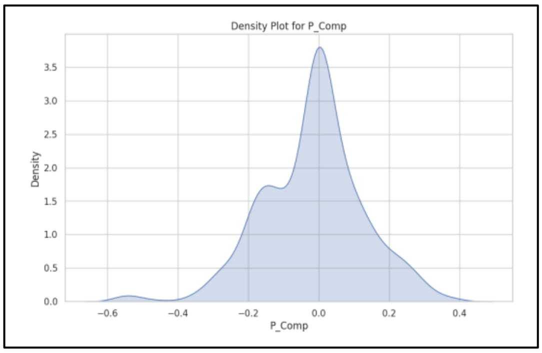


Fig. 8. Density Plot for P_Comp

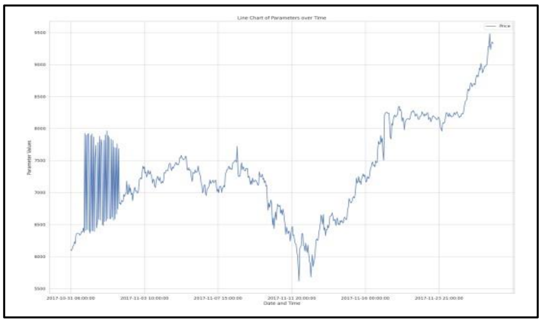


Fig. 9. Bitcoin Price Trend Over Time

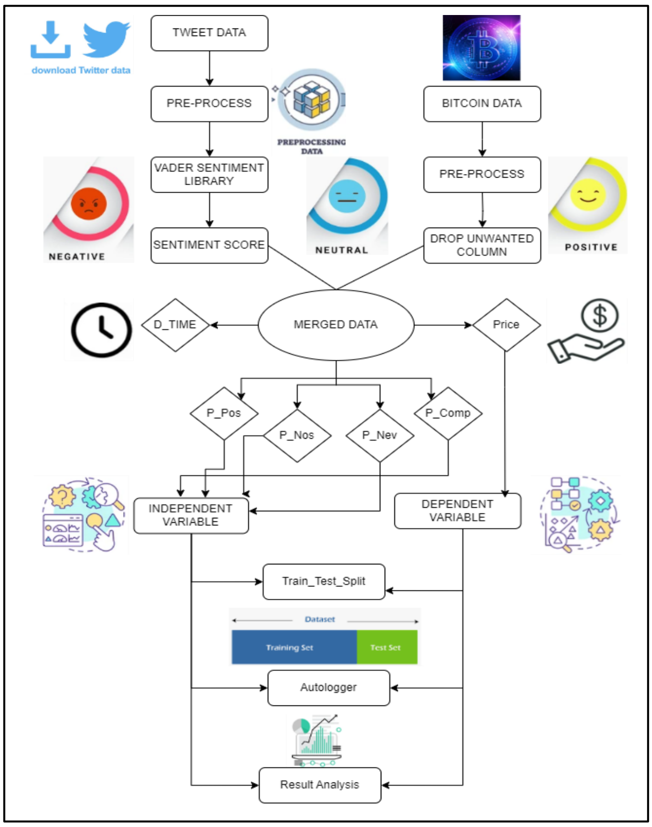


Fig. 10. Process Workflow

Methodology

The data was divided into training and testing sets with 80% and 20 % ratio respectively. This provided us with enough data to train our model and enough to test and evaluate over model.

The Vader sentiment library was used to get sentiment scores for the tweet data. Accordingly, we used scikit-learn to split of data and fit our data to the model. The methodology is presented in Fig. 10.

We additionally used the AutoLogger_ML library to train our data over various models at the same time. The results have been discussed in detail in the ‘Results’ section.

Results

After Training Data with AutoLogger_ML, the following were the results acquired.

TABLE II. **Results**

model	**training-mape**	**idation- use**	**lidation-mape**
linear regression	0.081905	2.52E+08	0.307097
and regression	0.082063	588230.9	**0.084925**
ridge regression	0.082174	587474	**0.084929**
elastic net	0.082152	589728.4	0.085083
decision tree regression	0.028081	913672.3	0.101768
random forest regression	0.050259	633763.3	0.088994
AdaBoost regression	0.083921	604257	0.089437
gradient boost regression	0.061982	586779.5	0.085896
xgboost regression	0.030267	684455.5	0.090393

The results for various models are shown in the previous table.

We achieved a minimum MAPE value of nearly 8.49% for used regression and ridge regression, which is less than the results achieved for the same dataset in [7].

Following Fig. 11 presents a comparison of the actual price vs the predicted price.

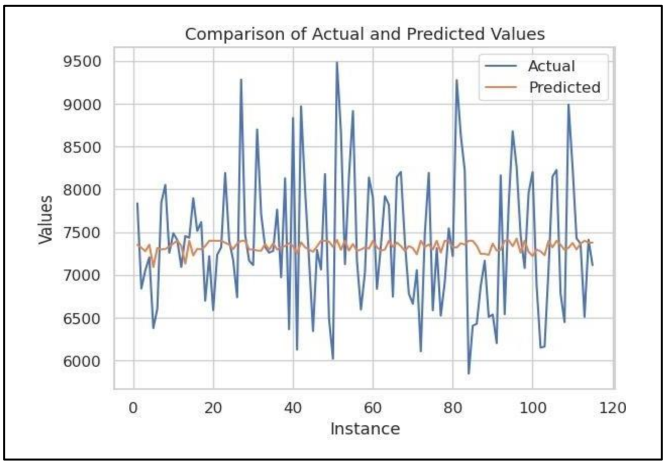


Fig. 11. Actual Vs Predicted

Conclusion

From this work, we concluded that the prices of bitcoin are very fickle and dynamic. Although tweets from authorized and powerful accounts have proven to impact the price of such cryptocurrencies. Using tweeter sentiment analysis we found some correlation between the values of the bitcoin and the sentiments of the tweets. Also, we were able to predict the value of the bitcoins using the sentiments score with a Validation-MAPE value of 8.45% with SGD- -regression and ridge-regression turning out to be the best models out of all tested.

References

[1] Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis - S M Raju, Ali Mohammad Tarif

[2] Bitcoin Price Prediction Considering Sentiment Analysis on Twitter and Google News - Ameni Youssfi Nouira, Mariam Bouchakwa, Yassine Jamoussi

[3] Improving the Prediction Accuracy in Deep Learning-based Cryptocurrency Price Prediction - Furkan BALCI.

[4] Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model - Ahmed Ibrahim.

[5] Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning - Franco Valencia, Alfonso Gómez- Espinosa, Benjamín Valdés-Aguirre

[6] BPTE: Bitcoin Price Prediction and Trend Examination using Twitter Sentiment Analysis - Muhammad K. Shahzad; Laiba Bukhari; Tayyeba Muhammad Khan; S. M. Riazul Islam; Mahmud Hossain; Kyung-Sup Kwak

[7] Cryptocurrency price prediction using Twitter sentiment analysis - Haritha G B and Sahana N B

[8] BPTE: Bitcoin Price Prediction and Trend Examination using Twitter Sentiment Analysis - Muhammad K. Shahzad; Laiba Bukhari; Tayyeba Muhammad Khan; S. M. Riazul Islam; Mahmud Hossain; Kyung-Sup Kwak

[9] Prediction of Cryptocurrency Price using Machine Learning Techniques and Public Sentiment Analysis - Mehedi Hasan Mishal, Nura Jannat Rakhi, Fahmida Rashid, Kawsar Hamid, Md Kishor Morol, Abdullah Al Jubair, Dip Nandi

[10] Predicting Stock Returns Using Graph Convolutional Networks - Zheng, Siyuan, et al

[11] Predicting Stock Market Price: A Logical Strategy using Deep Learning - Milon Biswas; Atanu Shome; Md. Ashraful Islam; Arafat Jahan Nova; Shamim Ahmed

[12] A Novel Method of Blockchain Cryptocurrency Price Prediction Using Fractional Grey Model - Yunfei Yang, Jiamei Xiong 2, Lei Zhao, Xiaomei Wang, Lianlian Hua and Lifeng Wu

[13] Predicting Stock Prices Using Natural Language Processing - Yi Zhang, Xu Han, and Jiawei Han

[14] Cryptocurrency Price Prediction Using LSTM and GRU Networks -Andrei-Alexandru Encean; Daniel Zinca

[15] Cryptocurrency Price Prediction Using Neural Networks and Deep Learning - Sumit Biswas; Mohandas Pawar; Sachin Badole; Nachiket

Galande; Sunil Rathod

[16] Predicting Cryptocurrency Prices Using Regression Techniques - R. Kavitha and Dr. K. R. Venugopal

[17] Kale, Sunil D., et al. "A comprehensive review of sentiment analysis on Indian regional languages: Techniques, challenges, and trends." International Journal on Recent and Innovation Trends in Computing and Communication 11.9s (2023): 93-110

[18] Bhagat, Chaitanya, and Deepak Mane. "Text categorization using sentiment analysis." Proceeding of International Conference on Computational Science and Applications: ICCSA 2019. Springer Singapore, 2020.