Twitter Sentiment Analysis for Bitcoin Price Prediction

Achyut Jagini
Btech Computer Science
PES University
Bangalore
achyut.jagini@gmail.com

Kaushal Mahajan
Btech Computer Science
PES University
Bangalore
kaushal.pes@gmail.com

Namita Aluvathingal
Btech Computer Science
PES University
Bangalore
namita0201@gmail.com

Vedanth Mohan
Btech Computer Science
PES University
Bangalore
vedanth.pes@gmail.com

Prajwala TR
Associate Professor
PES University
Bangalore
prajwalatr@pes.edu

***Abstract*—**

Cryptocurrencies, like Bitcoin, have become increasingly popular over the last decade. The price of Bitcoin has gone through several cycles of highs and lows. As a result, it is a widely discussed topic, especially on platforms like Twitter.

Sentiment analysis is a research area of Natural Language Processing. It is used to determine whether the text is positive, negative, or neutral. Twitter tweets are more challenging to analyze when compared to other forms of text, due to the presence of irregular grammar, emoticons, and sarcasm.

This project aims to analyze the effect of tweets on the stock price of Bitcoin. In order to study the effect, the sentiment associated with each tweet is calculated using VADER, and also the profession and follower count associated with verified users who tweet about bitcoin is found. Following this, a model is trained and tested using a combined dataset of tweet related data and historical bitcoin price data . It was found that the sentiment of tweets does correlate with the shift in the price of bitcoin.

Keywords—Bitcoin, Sentiment Analysis, Valence Aware Dictionary and sEntiment Reasoner, Twitter, Linear regression

I. INTRODUCTION

Cryptocurrency has gained a lot of momentum over the past decade. Bitcoin is one such cryptocurrency developed by Satoshi Nakamoto. It has a decentralized existence and is not regulated by any government. The price of Bitcoin constantly fluctuates in real-time.

Twitter is a social network site on which users interact through tweets and replies. It is used by users from different parts of the world and with different professions to speak about matters they feel passionately about. Fluctuations regarding cryptocurrency prices are often addressed on social media and are talked about by influencers and commoners alike. Users tweet about their predictions and other points of interest with regard to Bitcoin.

A person wishing to sell or buy Bitcoin searches for 'bitcoin' in the Twitter search bar and looks for tweets that relate to Bitcoin which may assist in predicting its price or value in the future. They would tend to trust people with influence in the market or experience in the field.

We perform sentiment analysis on tweets relating to Bitcoin to predict its price fluctuations. This could help those interested in investing gain a better perspective on when it would be a good time to invest.

In this project, the model is restricted to only Bitcoin as it is the most established in both market share and age.

Twitter tweets are obtained for a period of 14 months from February 2021 to April 2022.

2. LITERATURE REVIEW

D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel, and B. K. Lama (2018) collected tweets regarding Bitcoin from various sources and categorized them as positive or negative. For text categorization, they used Word2Vector and Bag-of-Words modes. The results from these techniques were processed with five distinct modes:
Multinomial, Naïve Bayes, Linear Support Vector Classifier, Bernoulli Naïve Bayes, and Random Forest. A classifier, which takes the output from each of the five modes as input and then categorizes the new tweet to the class with the maximum vote, was made. The cumulative sentiment was found for each day. The acquired positive and negative sentiments were fed to an RNN model alongside historical prices to forecast the new cost for the next time frame. The correlation of sentiments with prices was done using the Pearson Correlation test. The accuracy of sentiment categorization of tweets into two classes is 81.39% and the overall precision using the RNN mode is 77.62%.[3]

E. Şaşmaz and F. B. Tek (2021) targeted NEO altcoin and collected and filtered tweets that contained NEO in the hashtags by directly scraping them from Twitter. The period over which tweets were considered was five years. This data was then classified manually followed by feeding it as the

input to a random forest mode. The second phase of the project included investigating if the results of the daily sentiment had a reaction to the fluctuation in NEO's price. There was a positive correlation between the two. It was assumed and later found that BTC and Ethereum affect the prices of the cryptocurrencies and therefore even Bitcoin and Ethereum tweets are collected along with NEO tweets. The daily prices in Dollars and transaction volume of BTC and Ethereum were collected from Yahoo Finance. Python Scikit Learn library and The GridSearchCv were used to train the sentiment analyzer and the 'CountVectorized' method was used to change tweets to token counts having parameters. The results obtained were then compared with BERT Mode.[5]

Otabek Sattarov, Heung Seok Jeon, Ryumduck Oh, and Jun Dong Lee (2020) collected tweets from Twitter, Reddit, and Bitcointak.org over 60 days and performed sentiment analysis using VADER. The reaction between tweet sentiments and prices was analyzed. Random Forest was used with various features as inputs and outputs were analyzed. The error was measured as the difference between each model's prediction result with the closing Bitcoin cost and taking an absolute value. The average accuracy was 62.48%. In the prediction stage, the model had lost beyond 10,000 data points which could have provided better performance had they been included. We could confirm the correlation between tweet sentiments and Bitcoin prices but wanted better accuracy.[7]

Abdu Rehman Khurshid (2021) investigated the impact of social media and other sources of information to anticipate cost changes for two cryptocurrencies: Bitcoin and Cardano. Sentiment analysis was done using VADER. Both volumes of tweets and Google Trends were found to be correlated with the cost. inputs to the mode were sentiment analysis of collected Bitcoin and Cardino tweets, Google Trends data, and tweet volume. By utilizing Google trends, the prevalence of digital currency throughout recent years could be extracted, this information was used for prediction. Linear regression was applied to calculate the daily closing price of Bitcoin. A model created by utilizing Neural networks NN, SVM, and random forest RF was used and showed that predicting prices is feasible through analysis of sentiment and machine learning tools. The LSTM model was also used. The required data was taken from SinaWeibo, a Chinese social media platform. LSTM coupled with the historical cryptocurrency prices was used to predict future prices. The model had an accuracy of 87%. Multiple modes were attempted to figure out the superiority.[9]

Sara Abdai and Ben Hoskins (2021) used about 272,304 tweets from a dataset on Kaggle. They took tweets mentioning Bitcoin and aggregated them into one-minute "buckets" to use as input. To extract features from these buckets, they bundled the tweets in each bucket together, tokenized them, and removed punctuation, emojis, URLs, and stopwords. They then used one hot encoding and BERT to extract the features. They used a historical bitcoin price dataset from Finance public dataset and used 'Open Time' UTC timestamp denoting the time at the beginning of each one-minute interval) and 'Close' (the price of Bitcoin in US dollars at the end of that interval) in the mode. They used these changes in price to assign labels to each set of tweets.
They used Naive Bayes and SVM (two separate modes) to display a prediction of whether the cost would go higher or lower over a day. They compared these modes to a logistic regression model that used features generated by feeding Twitter data into the BERT mode. They achieved a training accuracy of 78% and a peak accuracy of 63%. The results from their experiments showed that in training using 2 or 3 labels, SVM outperformed the baseline algorithm, Naive Bayes, and BERT.[10]

3. DATA

Two datasets have been used in the project.

- Bitcoin tweets dataset.
- Bitcoin prices dataset.

3.1Bitcoin tweets dataset:

The inferences about the dataset are:

- The dataset is from Kaggle, named "Bitcoin Tweets".
- The data was primarily scraped from Twitter using the Twitter API "Tweepy" on the following conditions:
  - Date -All tweets in the range of the 5th of February 2021 and the 26th of April 2022, a span of 436 days,
  - Tweets containing the hashtags "Bitcoin" or "btc" were gathered.
  - The tweet must be in English.
- The resulting dataset contains 28,30,476 tweets related to bitcoin along with 12 other attributes. Each observation represents an English tweet.

TABLE I. Bitcoin tweets dataset

	ATTRIBUTE	TYPE	VALUES	UNIQUE	NANS
0	user_name	object	2830476	443139	31
1	user_location	object	2830476	72344	1403546
2	user_description	object	2830476	440375	352680
3	user_created	object	2830476	430262	93
4	user_followers	object	2830476	67584	138
5	user_friends	object	2830476	28043	138
6	user_favourites	object	2830476	118726	138
7	user_verified	object	2830476	20	138
8	date	object	2830476	2208780	138
9	text	object	2830476	2765189	138
10	hashtags	object	2830476	643626	17183
11	source	object	2830476	2152	3754
12	is_retweet	object	2830476	2	441

The "text" column contains all the tweets.
The Columns- user_created, source, user_friends don't appear to have any significance and aren't very useful.

The user_location column is 50% NaNs and does not have any pattern in it.

3.2 Bitcoin prices dataset:

This dataset was obtained using Yahoo Finance API and contains the bitcoin prices (in US dollars) for the range of dates corresponding to the dates of tweets, i.e 5th of February 2021 to 26th of April 2022.

The dataset contains 5 columns “Open” , “High” , “Low” , “CloseAdj” , “Close” and “Volume” with the index being the date.

TABLE 2. Bitcoin price dataset

Date	Open	High	Low	Close	Adj Close	Volume
2021-02-06	38138.386719	40846.546875	38138.386719	39266.011719	39266.011719	71326033653
2021-02-07	39250.191406	39621.835938	37446.152344	38903.441406	38903.441406	65500641143
2021-02-08	38886.828125	46203.929688	38076.324219	46196.464844	46196.464844	101467222687
2021-02-09	46184.992188	48003.722656	45166.960938	46481.105469	46481.105469	91809846886
2021-02-10	46469.761719	47145.566406	43881.152344	44918.183594	44918.183594	87301089896

4. METHODOLOGY

4.1 Preprocessing datasets

Preprocessing is done on the tweets and prices datasets by applying the following steps.

1)Bitcoin tweets dataset:

- Rows where “text” is NaN are removed.
- Drop “is_retweet” because it contains only "nan” or “false".
- Dropping the columns “hashtags”, ”source”, ”user_favourites”, and “user_friends” columns are removed.
- The user_location is also dropped as it is almost 50% NaNs, is very random and can’t be used in any way.
- There are a few observations where “user_verified” is neither True nor False. These rows are deleted.
- All duplicate rows were deleted.
- Reducing “date” which was the UTC timestamp to only date.
- Rows with NAN values are removed. Rows, where the “user_verified” value is neither True nor False, are deleted.
- UTC timestamp is reduced to only the date value.

The processed dataset now has 28,30,323 observations with the attributes:- ”user_name”, “user_description”, “user_followers”, “user_verified”, “date” and “text”.

2)Bitcoin price dataset:

- Drop column “Adj Close”.

4.2 Preprocessing Tweets and getting their sentiment score

4.2.1 Generating a slang dictionary

Social media platforms overflow with slang. These words are not part of the regular English language and therefore it is difficult for a sentiment analyzer to find the sentiment of these words.

In an attempt to overcome this a Slang dictionary was created by scraping slang along with its meaning from a couple of websites using Selenium.

This dictionary contains 1500+ slang.

4.2.2 Preprocessing Tweets

Tweets are a combination of expressions, emoticons, slang, symbols, URLs, and user's mentions. This is because of the casual nature of social media use by people. Raw tweets contain a lot of noise and can’t be fed directly to the sentiment analyzer due to this.

Therefore these raw tweets must be pre-processed in order to get a more accurate result.

The pre-processing techniques used on tweets are as follows:

- Replacing “\n” to space
- Removing mentions, hashtags, and links.
  Hashtags are removed by iterating over the tweet text and removing words starting with #symbol.
- Converting slang to its full form. The slangs are converting to full form by iterating over the text and expanding if present in the slang dictionary.
- Lemmatizing the tweet.The lemmatization is done on the tweet text by tokenization and lemmatizating the tokenized words using python natural language toolkit(NLTK) library WordNetLemmatizer.

4.2.3 Sentiment analysis

VADER or Valence Aware Dictionary and sEntiment Reasoner is a lexicon and rule-based tool used for sentiment analysis and is specifically attuned to sentiments expressed in social media, it also assigned scores for emoticons and thus it was a perfect match for the project. VADER has also been used by some of the researchers who have worked on this topic, such as Evita Stenqvist and Jacob Lönnö.[8]

VADER also not only returns the polarity of the string but also the magnitude of the polarity. It returns 4 sentiment fields:

1. Negative(0 to 1) 2.Neutral(0 to 1)
3.Positive(0 to 1) 4. Compound(-1 to 1)

The processed tweets are fed to VADER and the Compound score was chosen to represent the sentiment score of the tweet as it takes the other 3 scores into account while being calculated.

TABLE 3 Preprocessing tweets results

Original Tweet	Best case fr #Bitcoin, as the currency of the future I've ever listened to. #AustrianSchoolOfEconomics\n https://t.co/oLV3ue9gIm\n@TonyMu rega @MwangoCapital @MihrThakar @cheruiyotkb
Processed Tweet	Best case for real Bitcoin, as the currency of the future I've ever listened to.
VADER Score	{'neg': 0.0, 'neu': 0.741, 'pos': 0.259, 'compound': 0.6369}
Sentiment Score	0.6369

4.3 Obtaining a Tweet score

Just the sentiment and volume of tweets are not enough to predict the price of bitcoin, therefore the number of followers of the user and the profession of the user are also considered to obtain a "tweet_score" after all the impact of a tweet is only as good as its reach and credibility.

4.3.1 User Profession

The credibility of a tweet is not only determined by its content but also by the user and their credibility or expertise on that particular subject. One way to confirm their expertise on the matter is to check their background or profession and if it is related to the topic.

For bitcoin, we have identified a few prominent professionals that have knowledge of bitcoin.

```
["Financial Analyst","Journalist","Research Analyst",\
 "Investment Analyst","Cryptocurrency Analyst","Economist"\
 "Blockchain security architect","crypto security architect"\
 ,"Blockchain Developer","Mining technician"\
 ,"Consultant","Trader","Software Engineer","Blockchain"]
```

FIG 1 Bitcoin-related professions

The profession of all users can't be found due to reasons such as lack of information available, common names, the account may be an information page or impersonation to name a few. Therefore we decided to limit gaining the profession to only the "verified" users.

Selenium was used to scrape the professions of "verified" users from Google.

Each user was provided a score based on their profession/professions -

A verified users' profession_score was incremented by 1 for each profession of the user that was in a list of professions related to bitcoin.

For a user who is not verified or whose profession isn't taken to be related the score is 0.

4.3.2 Calculating tweet score

The Final score of the tweet is obtained using the sentiment score of the tweet, users' followers, and the users' profession score.

Tweet_ Score =(tweet_sentiment_score)*(user_followers)* (profession_score+1) (1)

The Tweet score incorporates the reach factor using user_followers and the verified users' profession_score helps prove the credibility of the tweet.

Multiplication is used due to the fact that the tweets' sentiment would spread as a factor of the users' followers which is similar to how an actual tweet spreads.

4.4 Tweets and Cryptocurrency Prices

The data from the tweets dataset is aggregated and stored as

- Average tweet score
- Number of tweets or tweet volume

for each day.

TABLE 4. Tweet aggregates

Date	Avg_tweet_score	Tweet_vol
2021-02-05	592.361387	1694.0
2021-02-06	498.539178	3278.0
2021-02-07	219.357551	3030.0
2021-02-08	1328.579672	5647.0

This data is now combined with the daily price of bitcoin in the range of 2021-02-05 to 2022-04-26 with the Date as the key.

The combined dataset has features 'Open', 'High', 'Low', 'Tweet_volume', 'Avg_score', and 'Close', with Date as its index.

4.5 Model

Linear Regression is a machine learning algorithm that uses a supervised regression algorithm as its basis. Regression models target prediction values based on independent variables. It is deployed for finding out the relationship between variables and forecasting. It also takes into account the number of independent variables used.

The data is split into training and test data frames in a 70-30 split. 70% training data and 30% test data.

There are 305 samples in the training dataset and 131 samples in the testing dataset.

Linear regression is used as it can help identify the past trend

and can extrapolate it to the future based on the relationship between the multiple independent variables and the dependent variable needed to be predicted.

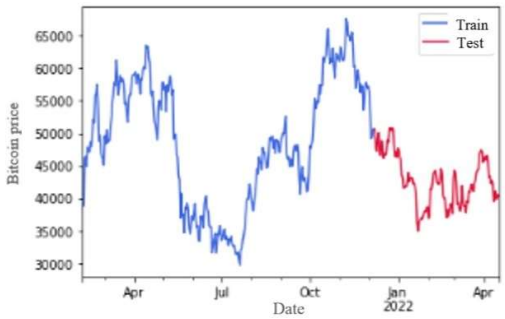

FIG 3. Actual closing price of bitcoin($)

The input features for the model are 'Open', 'High', 'Low', 'Tweet_volume', and 'Avg_score', and the Target feature is 'Close'.

The features and targets are standardized as they have varying ranges.

Predictions are obtained using Linear Regression and graphs are plotted.The model is validated using R^2-score and Mean Squared Error (MSE).

R^2-score(the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (2)$$

y_i-actual value

$\hat{y}_i$– predicted value

$\bar{y}$- mean of values

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

$$\text{MSE} = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (3)$$

y_i-actual value

$\hat{y}$ – predicted value

n-number of observations

5. RESULTS

The software used for coding the project is python version 3.8. Google colab is the code editor used.

The graphs are obtained for the predictions of closing price of bitcoin on the dataset.

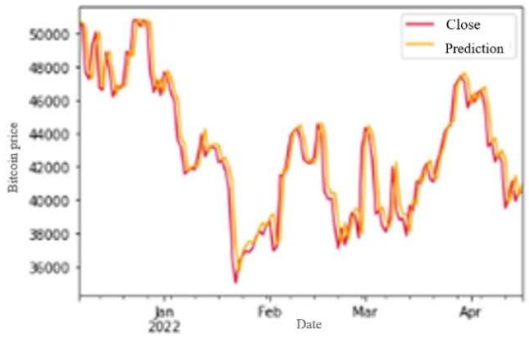

FIG 4. Actual price and predictions for price of bitcoin on test dataset($)

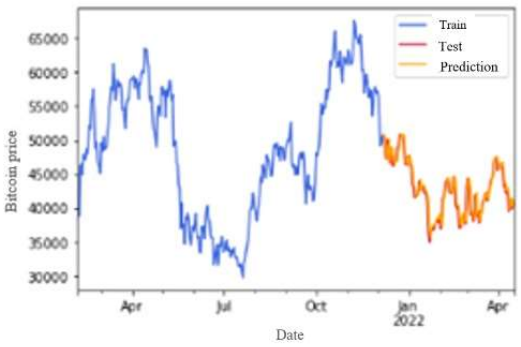

FIG. 5 Actual bitcoin price of train and test data with predictions on test data($)

TABLE 5. Statistics of prediction

TWEET AGGREGATES and PRICES	
Open_mean	48757.67644082992
High_mean	50116.59904585041
Low_mean	47200.239030481556
Tweet_volume_mean	17151.68797814208
Avg_score_mean	1494.0920273165386
Close_mean	[48795.80983607]
MSE	0.0038054912093926385
R2 score on Training set	99.31%
R2 score on Testing Accuracy	97.75%

As we can see in table 1, Linear Regression has a has MSE of 0.0038,a R^2-score of 99.31% or 0.9931 on the training data and an R^2-score of 97.75% or 0.9775 on the test data.

6.CONCLUSIONS AND FUTURE WORK

The goal of the project was to determine the effect tweets can have on Bitcoin prices and to do so rather than just factoring the sentiment of the tweet, we tried to assess the impact each tweet could have due to the tweeter in order to get a more just evaluation resulting in a better accuracy while eventually predicting the prices.

Linear regression was used and an R^2-score of 97.75% along with an MSE of 0.0038 was achieved.The aim of the project was to accurately predict future prices.

In order to make our model more accurate, future work could involve including more factors like the location, and tweets of all languages and incorporate others that affect Bitcoin prices as well. An attempt to include other cryptocurrencies as well can be made.

REFERENCES

[1]Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1 : No. 3, Article 1

[2]Critien, J.V., Gatt, A. & Ellul, J. Bitcoin price change and trend prediction through twitter sentiment and data volume. Finance Inov 8, 45 (2022). https://doi.org/10.1186/s40854-022-00352-7

[3]D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel and B. K. Lama, "Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), 2018, pp. 128-132, doi: 10.1109/CCCS.2018.8586824.

[4]E. Edgari, J. Thiojaya and N. N. Qomariyah, "The Impact of Twitter Sentiment Analysis on Bitcoin Price during COVID-19 with XGBoost," 2022 5th International Conference on Computing and Informatics (ICCI), 2022, pp. 337-342, doi: 10.1109/ICCI54321.2022.9756123

[5]E. Şaşmaz and F. B. Tek, "Tweet Sentiment Analysis for Cryptocurrencies," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 613-618, doi: 10.1109/UBMK52708.2021.955891

[6]Kilimci Z (2020) Sentiment analysis based direction prediction in bitcoin using deep learning algorithms and word embedding models. Int J Intell Syst Appl Eng 8:60–65

[7]O. Sattarov, H. S. Jeon, R. Oh and J. D. Lee, "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis," 2020 International Conference on Information Science and Communications Technologies (ICISCT), 2020, pp. 1-4, doi: 10.1109/ICISCT50599.2020.9351527.

[8] Evita Stenqvist and Jacob Lönnö. 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

[9]Khurshid, Abdul Rehman. "Cryptocurrency price prediction using sentiment analysis." Proceedings of conference. Washington, DC, USA. Vol. 17. 2017.

[10]Hoskins, Sara Abdali Ben. "Twitter Sentiment Analysis for Bitcoin Price Prediction"

[11]Suardi S, Rasel AR, Liu B (2022) On the predictive power of tweet sentiments and attention on bitcoin. Int Rev Econ Finance 79:289–301

[12] Valencia F, Gómez-Espinosa A, Valdes B (2019) Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy 21:1–12. https://doi.org/10.3390/e21060589

[13] Wołk K (2019) Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Syst 37. https://doi.org/10.1111/exsy.12493

[14] Zaman S, Yaqub U, Saleem T (2022) Analysis of bitcoin's price spike in context of Elon Musk's twitter activity. Glob Knowl Memory Commun

[15]Zhou X, Tao X, Yong J, Yang Z (2013) Sentiment analysis on tweets for social events, pp 557–562. https://doi.org/10.1109/CSCWD.2013.6581022