

# Dissertation on

# “Twitter Sentiment Analysis for Bitcoin Price Prediction”

*Submitted in partial fulfillment of the requirements for the award of degree of*

## Bachelor of Technology in

## Computer Science & Engineering UE19CS390B – Capstone Project Phase-2

***Submitted by :***

| **Achyut Jagini**  **Kaushal Mahajan**  **Namita Aluvathingal**  **Vedanth Mohan** | **PES2UG19CS013**  **PES2UG19CS178**  **PES2UG19CS245**  **PES2UG19CS449** |
| --- | --- |

*Under the guidance of*

**Dr. Prajwala TR**

### Associate professor

### PES University

### **June - Nov 2022**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

FACULTY OF ENGINEERING

**PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013) Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India



## PES UNIVERSITY

## (Established under Karnataka Act No. 16 of 2013)

Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

**FACULTY OF ENGINEERING**

# CERTIFICATE

*This is to certify that the dissertation entitled*

**‘Twitter Sentiment Analysis for Bitcoin Price Prediction’**

*is a bonafide work carried out by*

| **Achyut Jagini**  **Kaushal Mahajan**  **Namita Aluvatingal**  **Vedanth Mohan** | **PES2UG19CS013**  **PES2UG19CS178**  **PES2UG19CS245**  **PES2UG19CS449** |
| --- | --- |

In partial fulfillment for the completion of seventh semester Capstone Project Phase - 2 (UE19CS390B) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period June 2022 – Nov. 2022. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7th semester academic requirements in respect of project work.

| Signature  Dr. Prajwala TR  Associate Professor | Signature  Dr. Sandesh B J  Chairperson | Signature  Dr. B K Keshavan  Dean of Faculty |
| --- | --- | --- |

**External Viva**

| **Name of the Examiners**  **1.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Signature with Date**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_    **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| --- | --- |

# DECLARATION

We hereby declare that the Capstone Project Phase - 2 entitled **“Twitter Sentiment Analysis for Bitcoin Price Prediction”** has been carried out by us under the guidance of Dr. Prajwala TR, Associate Professor and submitted in partial fulfillment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester June – Nov. 2022. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

| **PES2UG19CS013**  **PES2UG19CS178**  **PES2UG19CS245**  **PES2UG19CS449** | **Achyut Jagini**  **Kaushal Mahajan**  **Namita Aluvathingal**  **Vedanth Mohan** | **Achyut Jagini**  **Kaushal Mahajan**  **Namita Aluvathingal**  **Vedanth Mohan** |
| --- | --- | --- |

# ACKNOWLEDGEMENT

### We would like to express our gratitude to Dr. Prajwala TR, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE19CS390B Capstone Project Phase – 2.

We are grateful to the Capstone Project Coordinators, Dr. Sarasvathi V, Professor and Dr. Sudeepa Roy Dey, Associate Professor, for organizing, managing, and helping with the entire process.

### We take this opportunity to thank Dr. Sandesh B J, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support we have received from the department. We would also like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

We are deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice- Chancellor, PES University for providing us various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement we have received from our family and friends.

# 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 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.

Through this project we aimed to analyze the effect of tweets on the stock price of Bitcoin. In order to study the effect, we deduced the sentiment associated with each tweet using VADER, and also took into consideration the profession and follower count associated with verified users who tweet about bitcoin. Following this we trained and tested our model using historical bitcoin price data. It was found that the sentiment of tweets does correlate with the shift in the price of bitcoin.

# TABLE OF CONTENTS

| **Chapter No. TITLE** | **Page No.** |
| --- | --- |
| 1. **INTRODUCTION**    1. Purpose Of The Project    2. Scope 2. **PROBLEM STATEMENT** 3. **LITERATURE REVIEW**    1. Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis    2. Recurrent Neural Network-Based Bitcoin Price Prediction by Twitter Sentiment Analysis    3. Twitter Sentiment Analysis for Bitcoin Price Prediction    4. Cryptocurrency Price Prediction using Sentiment Analysis    5. Tweet Sentiment Analysis for Cryptocurrencies    6. Summary 4. **PROJECT REQUIREMENTS SPECIFICATION**    1. Project Scope    2. Product Perspective       1. Product Features       2. User Classes and Characteristics       3. Operating Environment System       4. General Constraints, Assumptions and Dependencies       5. Risks    3. Functional Requirements    4. External Interface Requirements       1. User Interfaces 5. **DATA** 6. **SYSTEM REQUIREMENTS SPECIFICATION**    1. Hardware Requirements    2. Software Requirements    3. Functional Requirements    4. Non functional requirements    5. Advantages of system 7. **SYSTEM DESIGN**    1. Architecture Diagram    2. Use case diagram    3. Activity Diagram 8. **IMPLEMENTATION**     1. Cleaning the tweets Dataset    2. Create dataframe of slang data    3. Tweet Preprocessing    4. Extract professions of users       1. Scraping the users profession       2. Assigning Profession Score    5. Get Bitcoin price data    6. Combine datasets    7. Model    8. Model Validation 9. **RESULTS AND DISCUSSION** 10. **CONCLUSION** | **01**  01  02  **03**  **04**  04  05  07  09  10  11  **13**  13  13  14  15  **16**  **18**  18  18  18  18  19  **20**  20  21  22  **23**  23  23  23  25  26  27  27  29  **30**  **32** |
| **REFERENCES/BIBLIOGRAPHY**  **APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS**  **ANNEXURE-1**  **ANNEXURE-2** | **33**  **35**  **36**  **42** |

# LIST OF FIGURES

| **Figure No.** | **Title** | **Page No.** |
| --- | --- | --- |
| 7.1 | Architecture Diagram | 20 |
| 7.2 | Use Case Diagram | 21 |
| 7.3 | Activity Diagram | 22 |
| 8.1 | Preprocessing the dataset | 23 |
| 8.2 | Functions to process tweets | 24 |
| 8.3 | Cleaning and lemmatizing tweets | 24 |
| 8.4 | Code to obtain the profession of the user | 25 |
| 8.5 | List with bitcoin related professions | 26 |
| 8.6 | Generating bitcoin prices dataset | 26 |
| 8.7 | Calculating tweet score | 27 |
| 8.8 | Combining bitcoin price and tweet aggregates dataset | 27 |
| 8.9 | Model | 28 |
| 8.10 | Accuracy | 29 |
| 9.1 | Predictions for price of bitcoin | 30 |
| 9.2 | Bitcoin price of train and test data | 30 |
| 9.3 | Metrics | 31 |

# LIST OF TABLES

| **Table No** | **Title** | **Page No** |
| --- | --- | --- |
| 1.1 | Literature Survey | 11 |
| 2.1 | Tweet Dataset variables | 16 |

# 

# CHAPTER 1

**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.

## 1.1 Purpose of the project

This project aims to develop a model that can predict the price of bitcoin at a reasonable level of accuracy using the sentiment of bitcoin related tweets and the historical bitcoin price data.

## 1.2 Scope

* This is beneficial to those interested in investing as it gives them an idea of how the cryptocurrency will perform. Opinions are gathered from a social media site -Twitter- because it has established itself as one of the largest sites in the micro-blogging sphere and along with this also supports data collection.
* We chose to restrict our model to only bitcoin currently as it is the most established in both market share and age.

# CHAPTER 2

**PROBLEM DEFINITION**

Bitcoin is a [cryptocurrency](https://www.investopedia.com/terms/c/cryptocurrency.asp) that is not in the control of any specific person, organization, or entity, and thus removes the need for any external involvement during transactions. It can be bought, held, and sold for different purposes. People often want to know if investing in Bitcoin is a good decision and when is a good time to do so to maximize their gains and reap most benefits.

Bitcoin tweets have an effect on the price of bitcoin.

Our project aims to accurately find the fluctuation in Bitcoin price by performing sentiment analysis of tweets and also using historical bitcoin price data. This will help to serve the needs of those looking to understand or invest in Bitcoin themselves.

Our project follows the following broad steps

1. Collecting bitcoin related tweets and bitcoin historical price data for a certain time period.
2. Creating a model that can accurately find the fluctuation in the price of bitcoin with a reasonable degree of precision using the sentiment of bitcoin related tweets and using the historical bitcoin price data.
3. Finding accuracy of prediction.

# CHAPTER 3

**LITERATURE SURVEY**

## 3.1 “Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis”

**Otabek Sattarov**

# Collected tweets from Twitter, Reddit, and Bitcointalk.org for a period of 60 days starting from 12th March - 12th May 2018.

* Sentiment analysis was performed, using VADER. Before utilizing the typical results of sentiment analysis to figure out the heading of cost of Bitcoin, the connection between Tweet sentiment and price was determined. Random Forest was used with various features as inputs and outputs were analyzed.
* The error was measured by the difference (γ) (between each models prediction result(α) with the closing Bitcoin cost(β) and taking an absolute value: γ **= |**α-β|
* Module’s average accuracy was 62.48%.
* Module’s average accuracy error was 37.52 with the extremes being 43.83% and 21.84%.

**Advantages**

* A new model is used and findings of previous research and papers was used to decide on the models and approach.

**Limitations**

* In the prediction stage, the model had lost beyond 10000 data points which could have provided better performance.
* A sentiment lexicon developed specifically for Bitcoin would improve the correlation of the results of sentiment analysis on tweets and the Bitcoin’s variation in price, along with other variables such as hashtags, users, tweet volume, and emojis.

**Result**

* It is seen that there is a strong correlation of the Bitcoin price shift and Twitter sentiment.

## 3.2 “Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis”

**R Pant**

* The opinions on Twitter in regards to Bitcoin have an immediate or indirect influence on the general change in costs of Bitcoin. Research is concerned about forecasting the fluctuating cost of Bitcoin by investigating the opinion on Twitter to track down the connection between them.
* Tweets regarding Bitcoin collected from various sources are categorized as positive or negative.
* For text categorization the two models used were Word2Vector and Bag-of-Words.
* The acquired level of positive and negative tweets is fed to an RNN model alongside historical prices to foresee the new cost for the next time frame.
* New methodology of consolidating the sentiment scores along with historical cost to anticipate cost is applied.

1. Dataset: The gathered tweets are named physically as 'p' - positive, 'n' - negative and 'I' - unbiased or neutral. Absolute of 2585 ‘p’, 1669 ‘n’ and 3200 ‘I’ tweets are categorized manually.
2. Repeating and Irrelevant tweets were removed.
3. Regex and Weighted Search

**Sentiment Analysis**

The elements extracted from the two techniques for the 4,254 physically named tweets are then processed with five distinct models: Multinomial Naïve Bayes, Naïve Bayes, Linear Support Vector Classifier, Bernoulli Naïve Bayes and Random Forest.

* A classifier, which takes output from each of the five models as its input and then categorizes the new tweet to the class with maximum vote, is made.

Correlation of Sentiment with Price

* Pearson Correlation test used.
* The tweets from 01-01-2018 to 30-05-2018 were collected and the cumulative sentiment for each day was found.

**Advantages**

* The accuracy of prediction for sentiment categorization of tweets in two classes (p and n) is 81.39 % and the overall precision using the RNN model is 77.62%.

**Result**

* Word2Vector had an accuracy of 69.82% whole Bag-of-words had one of 78.50%.
* Bag-of-words is a better choice asWord2Vector does not perform well over a small dataset or at sentence level classification.
* The cumulative accuracy of prediction by classifier with a split in validation at 1:3 is found to be 81.4%.

## 

## 

## 

## 

## 

## 

## 3.3 Twitter Sentiment Analysis for Bitcoin Price Prediction

## Sara Abdali and Ben Hoskins

**Objective, Techniques/Methods**

* Measuring the sentiment of the average investor could be useful in making projections about the future movements of asset prices. Since the average sentiment of investors is not possible to measure, test whether the average sentiment of tweets about a specific asset could be used as a substitute.
* To improve the usefulness, focus on Bitcoin and its price movements. Take tweets mentioning Bitcoin and aggregate into one-minute "buckets" to use as an input. Use Naive Bayes and SVM (two separate models) to display a prediction whether the cost would go higher or lower over a day.
* Compared these models to a logistic regression model that used features generated by feeding twitter data into a language model called BERT.
* Combine the tweets already downloaded with a dataset from Kaggle. All tweets in this combined dataset mentioned Bitcoin in either a hashtag or in the main text of the tweet and were published between January and May of 2021. Total 272,304 tweets.
* Sort these tweets into "buckets" based upon the time they were published, segmenting the entire dataset into one-minute intervals. We were left with 34,294 buckets, each with a unique starting time.
* To extract features from these buckets, they bundled all the tweets in each bucket together. They tokenized the aggregate and removed punctuation, emojis, urls, and stopwords.
* Used one hot encoding and BERT to extract features.
* Bitcoin historical price data from Binance public dataset. Each entry in the dataset contained 12 data points, only using "Open time" and "Close" in the model.
* "Open time" is a UTC timestamp denoting the time at the beginning of each one-minute interval. Use these timestamps to align our labels with their corresponding tweet buckets. "Close" was the price of Bitcoin in US dollars at the end of that interval, which we could then use to find the ensuing change in price.
* Used these changes in price to assign labels to each of the sets of tweets. Performed experiments using two different label methods. The first assumed sets of tweets had either a “bullish” (positive) sentiment or a "bearish" (negative). Set the label to 1 if the price increased in the day following the tweets and -1 if the price decreased.

**Naive Bayes**

* splitting data into train, val, and test sets, used the sklearn implementation of the Naive Bayes model for training. Tuned the hyper- parameter alpha (Laplace smoothing parameter) for the model. This model had a training accuracy of 77% and a peak accuracy of 58% on the test set when alpha = 0.01.

**Support Vector Machines**

* use sklearn implementation of linear support vector classifiers for training. Fitting the data to support vector machines with radial and sigmoid kernels but they did not seem to work any better than a random classifier. As expected, the training accuracy keeps going up with increasing C, since the optimizer penalizes wrong classifications more strictly and over-fits to the training set more heavily.

**BERT +logistic regression**

* Used the vector representation of the CLS token as the feature for the model. Then fed these features into a logistic regression model, returning an accuracy of 58% with 2 labels and 50% with 3 labels on our test set.

**Advantages**

* After tuning the regularization parameter, C training accuracy of 78% was achieved and a peak accuracy of 63%. No on the test set when C = 0.01.

**Result**

* The result from experiments shows that in training using both 2 or 3 labels, SVM outperforms the baseline algorithm, Naive Bayes. Surprisingly, SVM also outperforms BERT. A possible reason behind this could be that a pre-trained BERT model is not familiar with financial terms and online slang used by cryptocurrency investors.
* Project has demonstrated that it is possible to predict Bitcoin price movements with relatively little data sourced solely from social media.

## 

## 

## 3.4 “Cryptocurrency Price Prediction using Sentiment Analysis”

**AR Khurshid**

* This paper investigates the impact of social media and other sources of information to anticipate cost changes for two cryptographic forms of money: Bitcoin and Cardano.
* Inputs to the model are sentiment analysis of collected Bitcoin and Cardino tweets along with Google Trends data and tweet volume.
* By utilizing Google trends, the prevalence of digital currency throughout recent years could be extracted, this information is utilized for the prediction.
* A sentimental analysis was performed and the information was dissected to decide whether it would be an important contribution to the final model.
* VADER was used and it determined tweets to be more neutral. Both volume of tweets and Google Trends were correlated with cost. Linear regression was applied to calculate the daily closing price of Bitcoin. Twitter’s sentiment on cryptocurrencies gravitates towards being positive irrespective of fluctuations in the cost.
* Social media sentiment and News were utilized to anticipate cryptographic prices. Bernoulli Naive Bayes, vector grouping, multinomial Naive Bayes were attempted.
* Model created by utilizing Neural networks (NN), SVM, random forest (RF) used for a selected few cryptocurrencies and it 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 Sina-Weibo, 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%

**Advantages**

* Multiple models have been used to figure out the superiority.
* All the models had a very high accuracy and provided useful insights.
* LSTM had a very high accuracy at 87%.

## 3.5 Tweet Sentiment Analysis for Cryptocurrencies

**E. Şaşmaz**

**Objective and Technique**

* Examined the applicability of sentiment analysis for cryptos. NEO altcoin was targeted and its data for the last 5 years was collected. All data containing “NEO” in their hashtag was collected and filtered. This data was then labeled/classified manually followed by feeding it as input for a random forest model.
* The second phase of the project included investigating if the results of the daily sentiment did have a relation with the fluctuation in NEO’s price. There was a positive correlation between the two.
* It is assumed that BTC and Ethereum affect the prices of all 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 was collected from Yahoo Finance.
* NEO tweets were scraped from Twitter.
* Python Scikit Learn library and The GridSearchCv are used to train the sentiment analyser and the CountVectorized method was used to change tweets to token counts having parameters. The results obtained were then compared with the BERT Model.

**Advantages**

* Details regarding crypto are widely spread across the internet, specifically on social media. Various social media personalities spread their thoughts and mindset on crypto and often use hashtags so it is very easy to gather data.

**Result**

1)Verified that the sentiment and price of NEO are correlated.  
2) The daily price of Bitcoin and ETH have an impact on the price of NEO.

## 

## 

## 3.6 Summary

| Paper Title | Author and year | Models used | Metrics |
| --- | --- | --- | --- |
| “Tweet Sentiment Analysis for Cryptocurrencies” | “E. Şaşmaz and F. B. Tek  2021” | Random Forest Classifier,BERT | 77%accuracy,45%accuracy |
| “Twitter Sentiment Analysis for Bitcoin Price Prediction” | “Sara Abdali,Ben Hoskins  2021” | * Naive Bayes and SVM * BERT | training accuracy of 78% and a peak accuracy of 63%. No on the test set when C = 0.01. |
| “Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis” | “D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel and B. K. Lama - 2018” | Text classification:  Word2Vector and Bag of-Words  Sentiment analysis:  Naïve Bayes, Bernouli Naïve Bayes, Multinomial Naïve Bayes, Linear Support Vector Classifier and Random Forest | Word2Vector has an accuracy of 69.8% and Bag-of-words is at 78.49%.  The tweet sentiment classification is 81.4% accurate and the RNN price prediction model is at 77.6%. |
| “Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis” | “Otabek Sattarov, Heung Seok Jeon, Ryumduck Oh, Jun Dong Lee-2020” | VADER,Random Forest Regression | 62.48%avg acc.  37.52 avg acc error with the extremes being 43.83% and 21.84% . |
| “Cryptocurrency Price Prediction using Sentiment Analysis” | “Abdul Rehman Khurshid  -2021” | VADER, Bernoulli Naive Bayes, multinomial Naive Bayes, and Linear support vector classification (classify news and social media sentiment)  NN, SVM and random forest for selected few cryptos, Averaged sentimental score of tweets over a period from 5 min.- 4 hours.  LSTM, historical cryptocurrency prices and data from Sina-Weibo | Logistic regression produced the best results.  LSTM 87 percent accuracy rate  aggregated method gave 83 % |

## 

# CHAPTER 4

**Project Requirement Specification**

## 4.1 Project Scope

Our project is meant to predict price fluctuations in Bitcoin price. This is beneficial to those interested in investing as it gives them an idea of how their crypto will do. Sentiments collected from a social networking site Twitter. We decided to limit the cryptocurrency to Bitcoin as it is the most used cryptocurrency and has the largest share.We do not, however, predict the actual price of Bitcoin overtime.

## 4.2 Product Perspective

Investing carries a certain degree of risk. While an investor might believe an asset's value will increase over time, when they purchase an asset, it is almost impossible to know whether they will gain or lose money on that investment. Being able to predict the future price movement of an asset is an extremely powerful tool for any investor.

### 4.2.1 Product Features

* Collects tweets using relevant hashtags and prices of bitcoin over a particular period of time
* Format this information and extract the important parts to perform sentiment analysis to obtain the emotion and classify them as positive or negative.
* Correlate this with the current price of bitcoin and predict the fluctuation in price.

### 4.2.2 User Classes and Characteristics

* **Investors**: Those interested in purchasing bitcoin as an investment. They would like to understand if the purchase is worth it by checking if it’s predicted to do well.
* **Analysts**: People interested in studying and understanding the trend in bitcoin price fluctuation. This also includes people using it in their projects and papers.
* **Media**: Newspapers, journals, etc, referring to the price fluctuation prediction to spread the word and further confirm what they think.

### 4.2.3 Operating Environment System

The Operating environment system will operate in Windows, Mac and Ubuntu operating systems using google colab.

### 4.2.4 General Constraints, Assumptions and Dependencies

* **Constraints and Dependencies**

1. Requires the tweets to have a hashtag to denote it’s about bitcoin/cryptocurrency. Some users may not tag their tweets.
2. Prediction is based solely off of tweets hence relies heavily on twitter.
3. Data repository and distribution requirements <https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets> <https://in.investing.com/crypto/bitcoin/historical-data>

* **Assumptions**

1. We assume that there is some correlation between tweets and price fluctuations.
2. Most tweets are tagged appropriately

### 4.2.5 Risks

* Sentimental analysis of tweets that include hashtags, emoticons or are sarcastic.
* Fewer tweets on particular days/weeks

## 4.3 Functional Requirements

* Predicting correlation of sentiment of tweets with bitcoin prices
* Predict the fluctuation in the price of bitcoin based on sentiments of tweets
* Make a graph of these fluctuations to better visually depict the changes

## 4.4 External Interface Requirements

### 

### 4.4.1 User Interfaces

Google Colab

### 4.4.2 Hardware Requirements

* Processor: 64-bit
* RAM: 4GB

### 4.4.3 Software Requirements

* Twitter API
* Python 3.7
* Libraries: Scikit-learn, Pandas, NumPy, Matplotlib and a few ML model libraries.
* Data on the prices of bitcoin for a certain time range. Sources: Kaggle, Google
* Tweets related to bitcoin during the same time range. Source: Twitter

# 

# 

# 

# 

# 

# 

# 

# 

# 

# CHAPTER 5

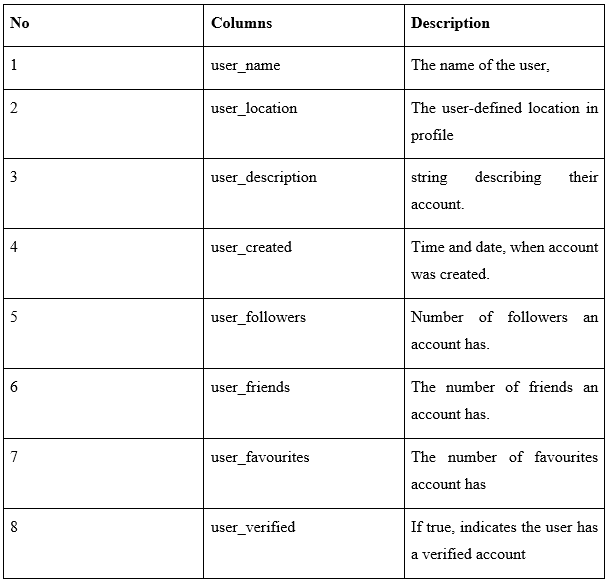
**DATA**

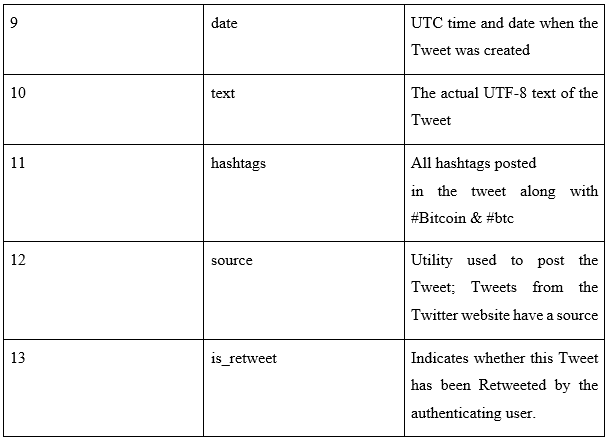
**Dataset**

For tweets, the dataset used is downloaded from kaggle

It is available on “https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets”. The tweets have #Bitcoin and #btc hashtag. The dataset consists of 13 columns.

**TABLE 2.1** Tweet Variables Dataset



****

The bitcoin historical data is downloaded from <https://in.investing.com/crypto/bitcoin/historical-data>.

The data has columns Date, Price, Open, High, Low, Volume.

# CHAPTER 6

**SYSTEM REQUIREMENTS SPECIFICATION**

## 6.1 Hardware Requirements

* Processor-64 bit
* Hard disk-80 GB
* RAM-4 GB

## 6.2 Software Requirements

* Operating System-Linux, Windows or MAC OS.
* Google colab
* Twitter API
* Python 3.7
* Libraries: Scikit Learn, Pandas, NumPy, Matplotlib and a few ML model libraries.

## 6.3 Functional Requirements

Functions to be performed by the system.

* Classify the tweets into positive, negative and neutral.
* Predict price of bitcoin

## 6.4 Non-functional requirements

* **Innovation**:Using a combination of historical prices, tweet sentiment and tweeter attributes to predict the fluctuation in price.
* **Performance**:System will try to provide results fast.
* **Security:** System is designed to be secure
* **Reliability:** The application has few chances of failure
* **Maintainability:** A well maintained code is possible with our project.
* **Portability:** Highly portable as it is being run on Google Colab, a cloud based service.
* **Legacy to modernisation:** Code developed using latest python libraries
* **Reusability:** Code follows good principles and is reusable
* **Compatibility:** Code is compatible on Linux, Windows and MAC OS.

## 6.5 Advantages of system

* High performance and gives results easily
* Ease of use
* Helps with prediction of fluctuation of Bitcoin price with good accuracy
* Can be used by companies and investors

# 

# 

# 

# 

# CHAPTER 7

**SYSTEM DESIGN**

## 

## 7.1 Architecture Diagram

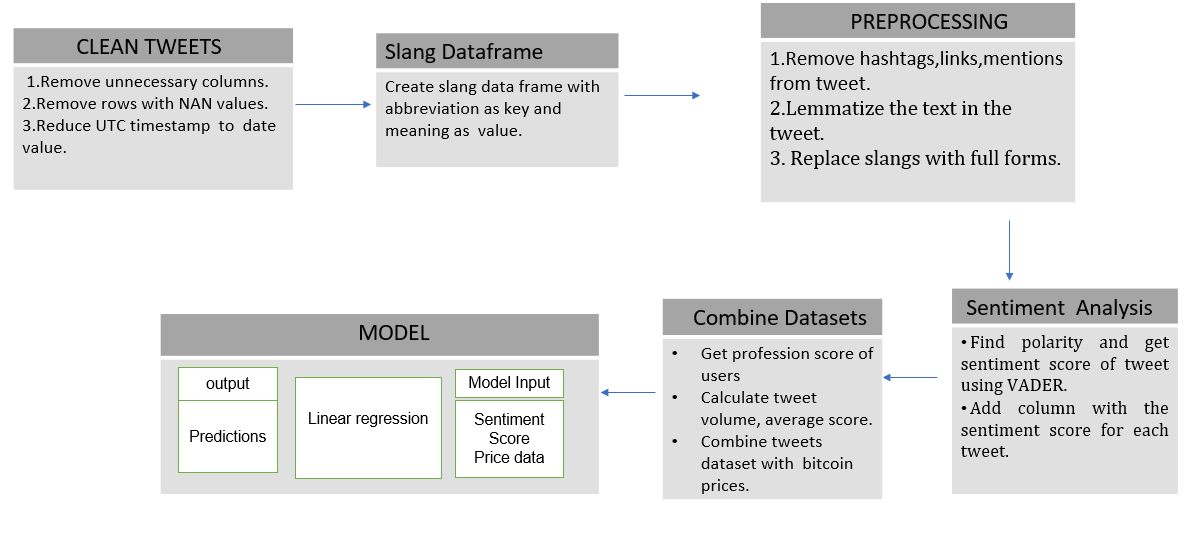


Fig 7.1-Architecture Diagram

## 

## 

## 

## 

## 

## 

## 7.2 Use case diagram

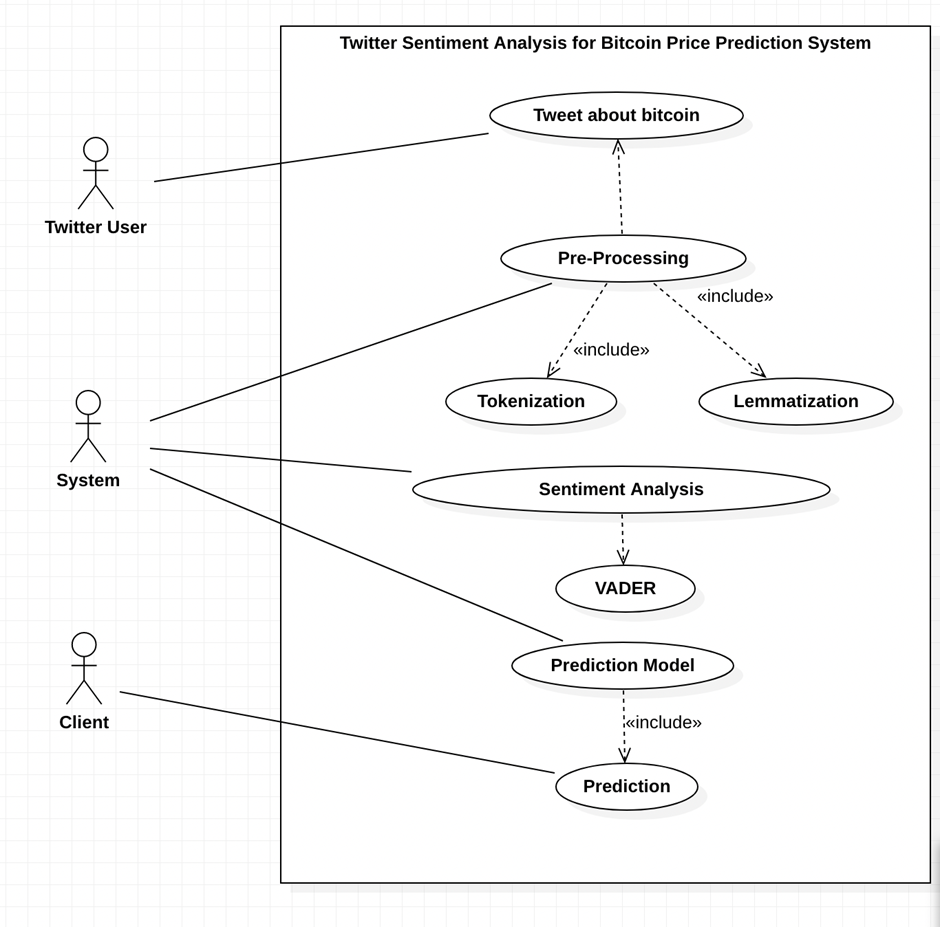
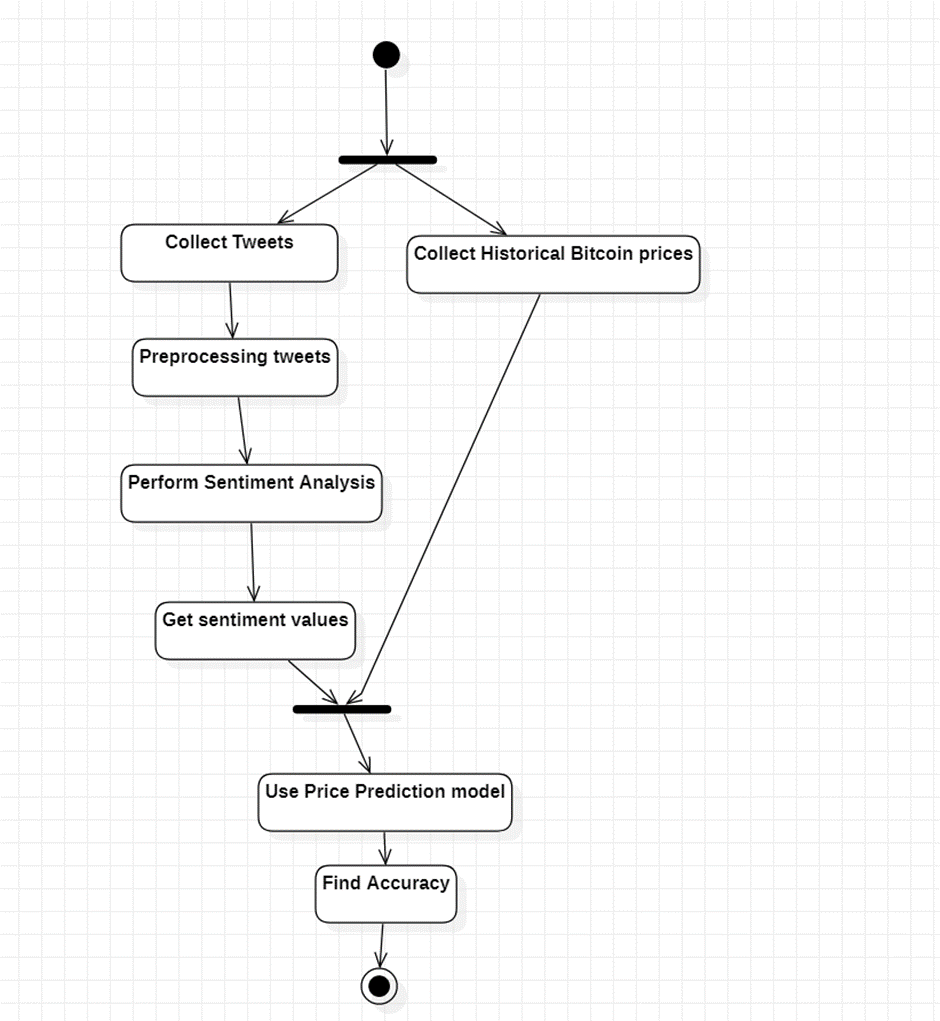


Fig 7.2-Use Case Diagram

## 

## 

## 7.3 Activity Diagram



7.3-Activity Diagram

# 

# CHAPTER 8

**IMPLEMENTATION**

## 

## 8.1 Cleaning the tweets Dataset

* First we must clean the dataset and get rid of unwanted values and attributes.

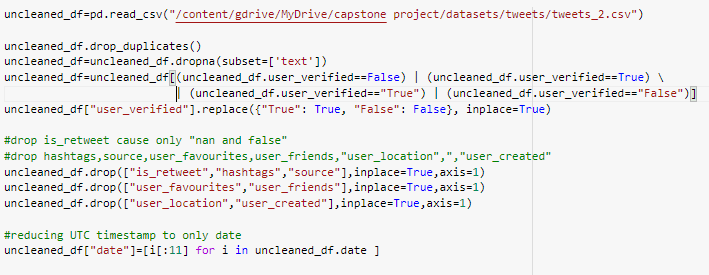


Fig 8.1-Preprocessing the dataset

## 8.2 Create dataframe of slang data

* Slangs in the tweets must be expanded to get a more accurate sentiment score while using vader.
* A slangdictionary is created by scraping 1500+ slangs from “<https://www.socialmediatoday.com/content/top-twitter-abbreviations-you-need-know>” and “<https://www.webopedia.com/reference/text-abbreviations/>” using BeautifulSoup.

## 8.3 Tweet Preprocessing

Hashtags, mentions and tags will hinder the accuracy of our sentiment score and thus they are removed. This cleaned text is checked for slangs and the slangs are expanded if present.

The text is now lemmatized and passed to VADER in order to obtain the sentiment score of the tweet.

****

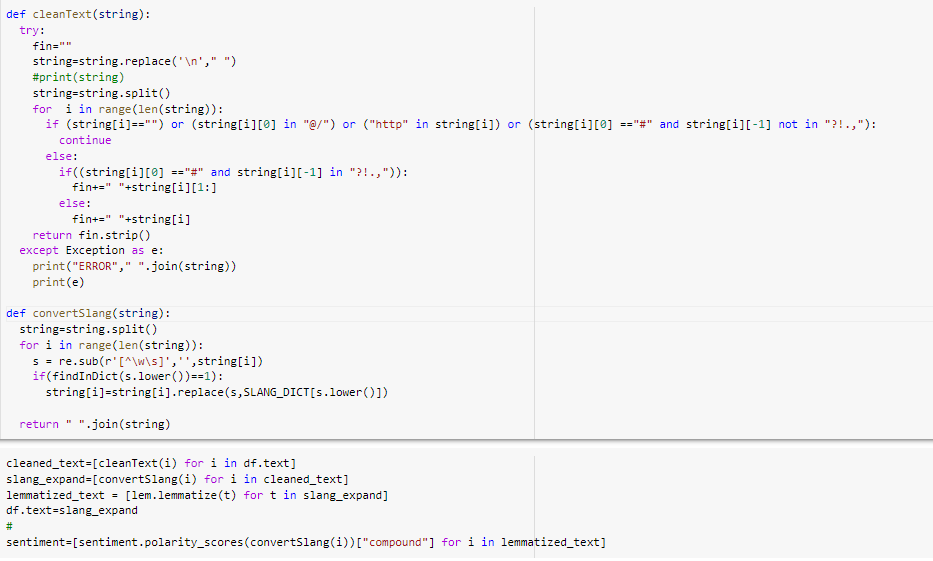
Fig 8.2-Functions to process tweets****

Fig 8.3-Cleaning and lemmatizing tweets

## 

## 

## 

## 

## 

## 

## 8.4 Extract professions of users

A tweet can only have an impact only if the user is trustworthy and for a tweet related to bitcoin, one way is to check if the user is “verified”. Another is to see if the profession of the user is related to bitcoin.

8.4.1 Scraping the users profession

* A list of all “verified” users if created and iterated through to get the professions of those users.
* Usernames are stripped of any emoticons and a query regarding their profession is passed to google.
* Selenium was used to scrape the resulting webpage and obtain the profession of the users.

8.4.2 Assigning Profession Score

* A “profession\_score” is assigned to each user.
* Initially 0 ,it is incremented if a/the profession of the user is present in a list of professions we believe is related to bitcoin.

The profession and the profession\_score are appended as columns.

****

Fig 8.4-Code to obtain the profession of the user

****

Fig 8.5-List with bitcoin related professions

## 

## 8.5 Get Bitcoin price data

* Generate a dataset containing the Date, Open price, Close price, High and Low for each day in the range of dates of the tweets.
* Yahoo Finance API was used to create this dataset.
* Save this dataset as a csv.

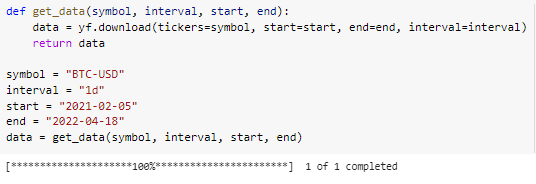


Fig 8.6-Generating bitcoin prices dataset

## 

## 8.6 Combine datasets

* A final score is calculated for each tweet using sentiment score of the tweet, the users’ followers and profession score.
* Calculate a tweet dataframe with tweet aggregates for each day.
* Merge BTC price data with twitter data using the date as the key.

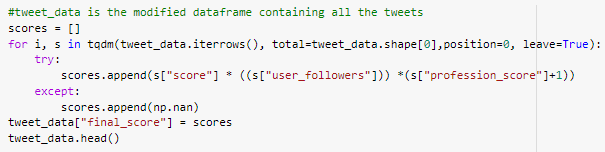


Fig 8.7-Calculating tweet score

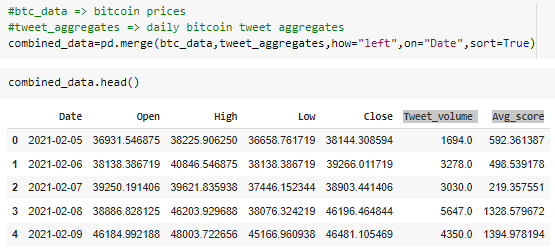


Fig 8.8-Combining bitcoin price and tweet aggregates dataset

# 

## 8.7 Model

* Data split into training and test dataframe 70% -train.
* Input features of model=['Open', 'High', 'Low', 'Volume', 'Tweet\_volume', 'Avg\_score']
* Target=['Close']
* The features and targets are standardized as they have varying ranges.
* Predictions obtained using Linear Regression and graph plotted.
* Model is validated using -score and RMSE .



Fig 8.9- Model

## 

## 

## 

## 8.8 Model Validation

R square score coupled with MSE is used to validate the model.

the Rsquare score and MSE is calculated for the training and testing set.

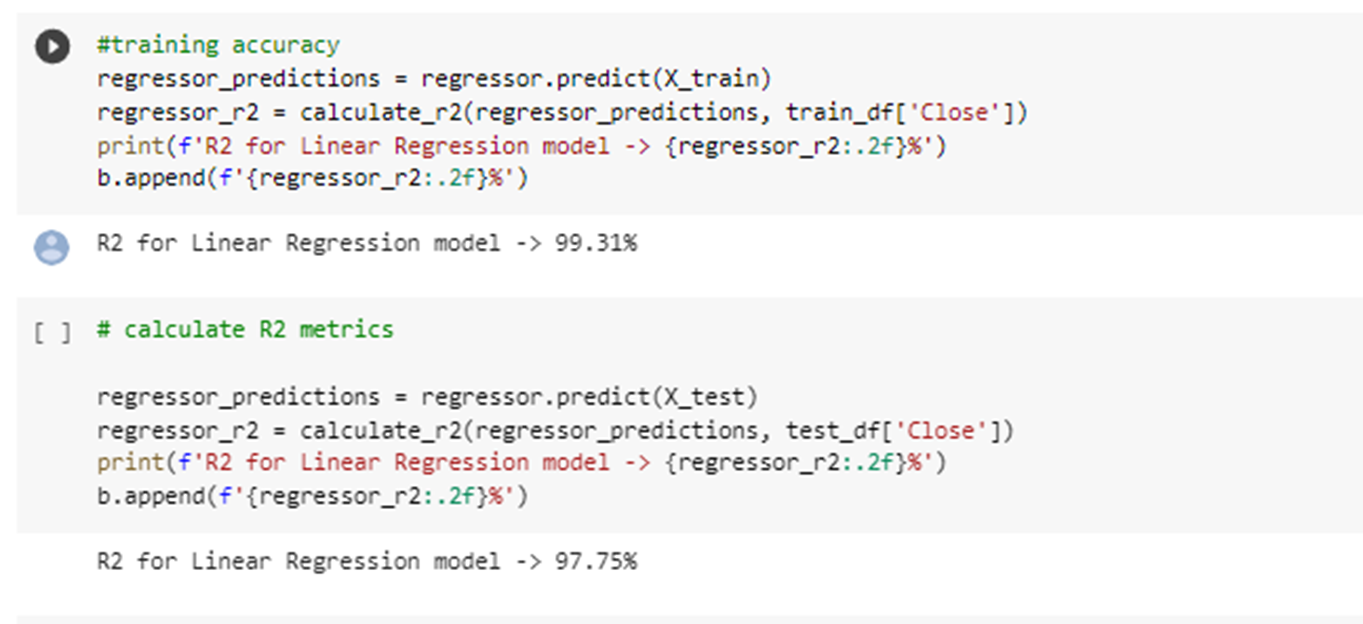


Fig 8.10- Accuracy calculation

# CHAPTER 9

**RESULTS AND DISCUSSION**

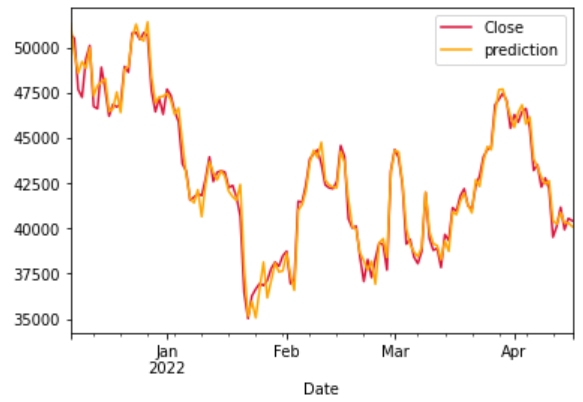


Fig 9.1- Predictions for price of bitcoin

In Fig 9.1, Orangeindicates the predictions for closing price while Crimson is the actual closing price.

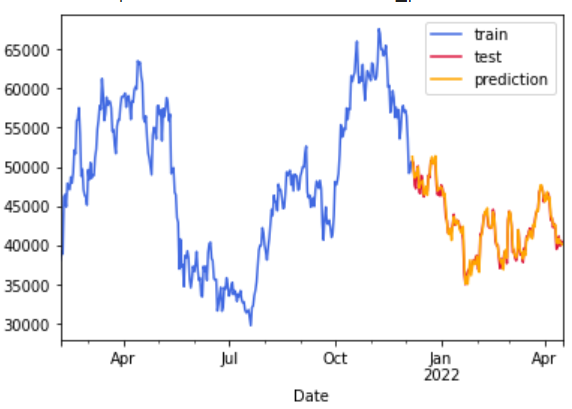
****

Fig 9.2- Bitcoin price of train and test data

In Fig 9.2, Red indicates the bitcoin price of training data and blue indicates bitcoin price of test data.Green indicates the predictions.

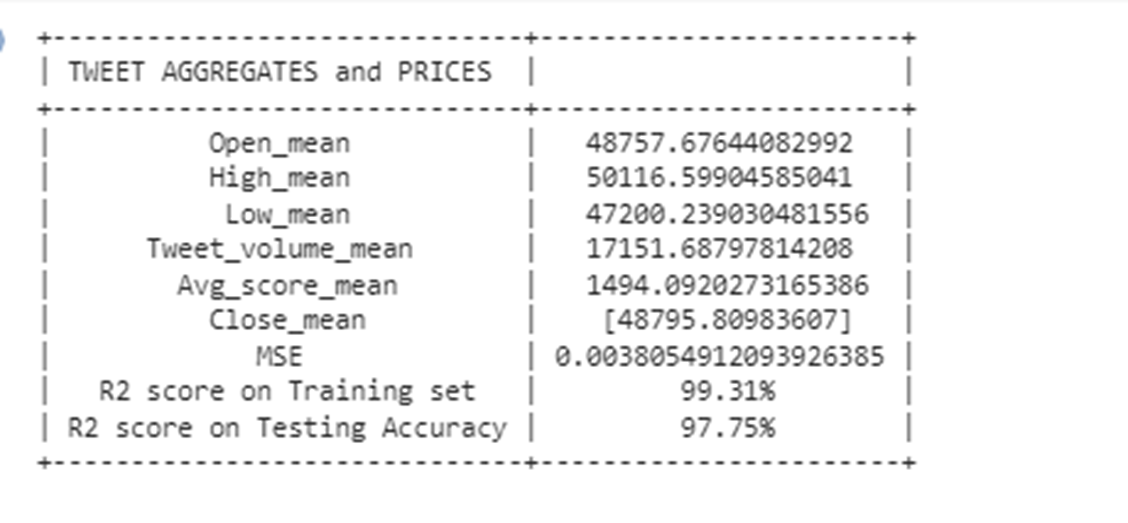


Fig 9.3- Metrics

The -score in the training data obtained is **99.31% or 0.9931**.

The -score of predictions obtained for the linear regression prediction model is found to be **97.75% or 0.9775**.

The Mean Square Error (MSE) for linear regression is found to be **0.003896**.

The results obtained are at par with initial estimates.

# 

# 

# 

# 

# 

# 

# CHAPTER 10

**CONCLUSION**

In the first phase of the project, the problem statement and scope of the project was defined. The architecture and dataset to be used was identified. By reading more about our topic and the papers published on it, we understood the model we can use, their advantages and limitations, gained better insight into the scope of the project, and validated our hypothesis.

In the second phase of the project ,the project was implemented. Our goal was to determine the effect tweets can have on Bitcoin prices so we could accurately predict future prices. The Bitcoin price was predicted and results obtained were at par with our assumptions and expectations.

In order to make our model more accurate, future work could involve including more factors that affect Bitcoin prices and we could also attempt to include other cryptocurrencies as well.

# REFERENCES/BIBLIOGRAPHY

[1]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>

[2]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.

[3]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

[4]Twitter Sentiment Analysis for Bitcoin Price Prediction - Sara Abdali Ben Hoskins 2021

[5]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.

[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]Kraaijeveld O, De Smedt J (2020) The predictive power of public twitter sentiment for forecasting cryptocurrency prices. J Int Finan Markets Inst Money 65:101188. <https://doi.org/10.1016/j.intfin.2020.101188>

[8]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

[9]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.

## 

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

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

[12]<https://in.investing.com/crypto/bitcoin/historical-data>.4

[13]<https://online.visual-paradigm.com/>

[14]<https://keep.lib.asu.edu/items/133184>

[15]<http://www.diva-portal.org/smash/get/diva2:1110776/FULLTEXT01.pdf>

[16]<https://helda.helsinki.fi/bitstream/handle/10138/344119/keningi_eino_thesis_2022.pdf>

[17]<https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets>

[18]https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

[19]https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/?ref=gcse

# APPENDIX

## APPENDIX A: Definitions, Acronyms and Abbreviations

**Definitions**

Tokenization -chunking down sensitive data to unique identifiers in such a way that they retain their original meaning.

Lemmatization - remove only inflectional endings and return the base form of a word

**Acronyms**

BTC -Bitcoin

ETH -Ethereum

RNN -Recurrent neural networks, a type of neural network class used to model sequential data.

VADER – “Valence Aware Dictionary and Sentiment Reasoner”, a sentiment analyzing tool that provides a sentiment score as well.

BERT – “Bidirectional Encoder Representations from Transformers”.

SVM- “Support Vector Machine” is a classification-based Machine Learning algorithm.

# ANNEXURE -I

Twitter Sentiment Analysis for Bitcoin Price Prediction

| Achyut Jagini Btech Computer Science PES UniversityBangalore [achyut.jagini@gmail.com](mailto:achyut.jagini@gmail.com) | Kaushal Mahajan  Btech Computer Science PES UniversityBangalorekaushal.pes@gmail.com | Namita Aluvithangal Btech Computer Science PES University Bangalore namita0201@gmail.com | Vedanth Mohan Btech Computer Science PES UniversityBangalore vedanth.pes@gmail.com | Dr.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 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.**

**Through this project we aimed to analyze the effect of tweets on the stock price of Bitcoin. In order to study the effect, we deduced the sentiment associated with each tweet using VADER, and also took into consideration the profession and follower count associated with verified users who tweet about bitcoin. Following this we trained and tested our model using 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, VADER, 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 we restrict our model to only bitcoin currently 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, Bernouli 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 were 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 mode was aso used. The required data was taken from SinaWeibo, a Chinese social media platform. LSTM couped with the historica 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 a 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 abes to each set of tweets. They used two different abe methods.

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**

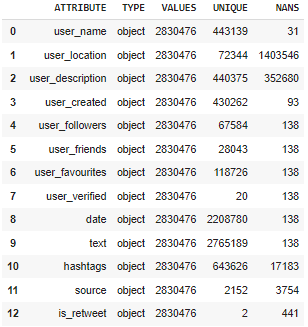
Two datasets have been used in the project.

* 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

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.

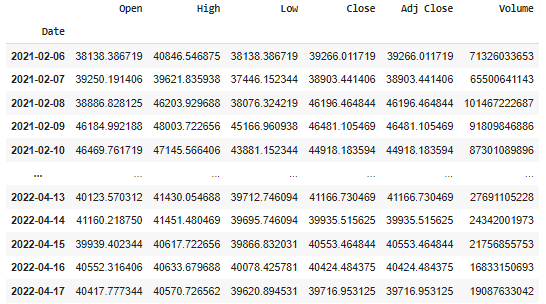
The user\_location column is 50%NaNs and does not have any pattern in it.

**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



**4. METHODOLOGY**

**4.1 Preprocessing datasets**

Data preprocessing is the process of transforming raw data into something that can be used by a machine learning model. It is the initial and most crucial phase in creating a machine-learning model.

*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 its sentiment score**

***4.2.1 Generating a slang dictionary***

Social media platforms overflow with slangs. 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 slangs along with their meaning from a couple of websites using Selenium.

This dictionary contains 1500+ slangs.

***4.2.2 Preprocessing Tweets***

Tweets are a combination of expressions, emoticons, slangs, 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.
* Converting slangs to their full form.
* Lemmatizing the tweet

***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@TonyMurega @MwangoCapital @MihrThakar @cheruiyotkb |
| --- | --- |
| Processed Tweet | Best case for real Bitcoin, as 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 |

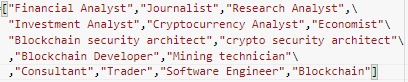
**4.3 Obtaining a Tweet score**

*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 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 professions that have knowledge on bitcoin.

******

**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 their 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 sentiment score of the tweet, users' followers and the users’ profession score.

Tweet\_ Score =

(tweet\_sentiment\_score)\*(user\_followers)\*(profession\_score+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 ofthe users’ followers which is similar to how an actual tweet spreads.

**4.4 Tweets and Cryptocurrency Prices**

The data from the tweets dataset 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 |

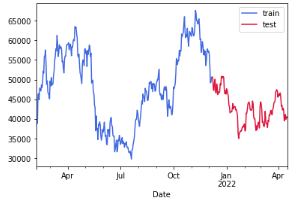
This data is now combined with the daily price of bitcoin in the range of 2021-02-05 to 2022-04-26 with the Date as the key.

The combined dataset has features 'Open', 'High', 'Low', 'Tweet\_volume', and '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 amount of independent variables used.

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



**FIG 2.** Closing for price of bitcoin($)

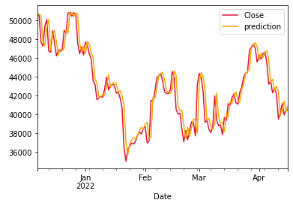
The input features for the model are 'Open', 'High', 'Low', 'Tweet\_volume', and 'Avg\_score' and the Target feature is ‘Close’.

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

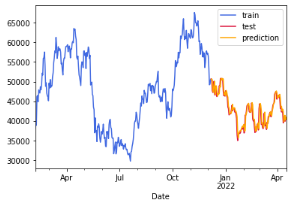
Predictions are obtained using Linear Regression and graph plotted.

Model is validated using R2-score and Mean Square Accuracy (MSE).

**5. RESULTS**

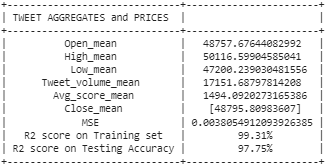
****

**FIG 3**. Predictions for price of bitcoin($)

****

**FIG. 4** Bitcoin price of train and test data($)

**TABLE 5.** Statistics of prediction



As we can see in table 1, Linear Regression has a has MSE of 0.0038 ,a -score of 99.31% or 0.9931 on the training data and an -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 of 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 the -score of 97.75% along with a 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, 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."

**ANNEXURE -II POSTER**

| **Group No:55** | **Title: Twitter Sentiment Analysis for Bitcoin Price Prediction**  **Domain:Machine Learning** | |
| --- | --- | --- |
| **Abstract:** Sentiment analysis is a research area of Natural Language Processing. It is used to determine whether a body of text is positive, negative, or neutral. Twitter tweets are technically more challenging to analyze when compared to other forms of text, due to the presence of irregular grammar, emoticons and sarcasm. In this project we aim to analyze various factors that might affect the stock price of Bitcoin, one such factor being tweets related to the popular cryptocurrency, ascertain their sentiment and try to predict the way in which the price will fluctuate. | | |
| **Team:**          **Supervisor:** | Achyut Jagini  PES2UG19CS013  Kaushal Mahajan  PES2UG19CS178  Namita Aluvathingal  PES2UG19CS245    Vedanth Mohan  PES2UG19CS449    Dr Prajwala TR | **Architecture/flow diagram** |