Dissertation on

"Twitter Sentiment Analysis for Bitcoin Price Prediction"

Submitted in partial fulfilment 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	**PES2UG19CS013**
Kaushal Mahajan	**PES2UG19CS178**
Namita Aluvathingal	**PES2UG19CS245**
Vedanth Mohan	**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	**PES2UG19CS013**
Kaushal Mahajan	**PES2UG19CS178**
Namita Aluvathingal	**PES2UG19CS245**
Vedanth Mohan	**PES2UG19CS449**

In partial fulfilment 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	Signature	Signature
Dr Prajwala TR	Dr. Sandesh B J	Dr. B K Keshavan
Associate Professor	Chairperson	Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. ____________________________	____________________________
2. ____________________________	____________________________

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 fulfilment 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	**Achyut Jagini**	**Achyut Jagini**
PES2UG19CS178	**Kaushal Mahajan**	**Kaushal Mahajan**
PES2UG19CS245	**Namita Aluvathingal**	**Namita Aluvathingal**
PES2UG19CS449	**Vedanth Mohan**	**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

LIST OF FIGURES

LIST OF TABLES

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 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: $\gamma = |\alpha-\beta|$
- 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 datapoints 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 are collected from various sources are categorised as positive or negative.
- 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 categorised manually.
2) Repeating and Irrelevant tweets were removed.
3) Regex and Weighted Search

- For text categorization the two models used were Word2Vector and Bag-of-Words are utilized.

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 it's input and then categorises 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% while 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

- Measure 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 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, bundled all the tweets in each bucket together. Tokenized the aggregated 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 used "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. We 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 our 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 our model. We 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 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 our baseline algorithm, Naive Bayes. Surprisingly, SVM also outperforms BERT. A possible reason behind this could be that a pretrained 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 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 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.
- LSTM model was also used. The required data was taken from a Sina-Weibo, a chinese social media platform. LSTM coupled with the historical cryptocurrency prices was used to predict future price. The model had an accuracy of 87 percent accuracy rate.

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 labelled/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 there 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 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 5min.- 4hours. 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 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.

No	Columns	Description
1	user_name	The name of the user,
2	user_location	The user-defined location in profile
3	user_description	string describing their account.
4	user_created	Time and date, when account was created.
5	user_followers	Number of followers an account has.
6	user_friends	The number of friends an account has.
7	user_favourites	The number of favourites account has
8	user_verified	If true, indicates the user has a verified account

9	date	UTC time and date when the Tweet was created
10	text	The actual UTF-8 text of the Tweet
11	hashtags	All hashtags posted in the tweet along with #Bitcoin & #btc
12	source	Utility used to post the Tweet; Tweets from the Twitter website have a source
13	is_retweet	Indicates whether this Tweet has been Retweeted by the authenticating user.

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: Scikitlearn, 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 and tweets 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

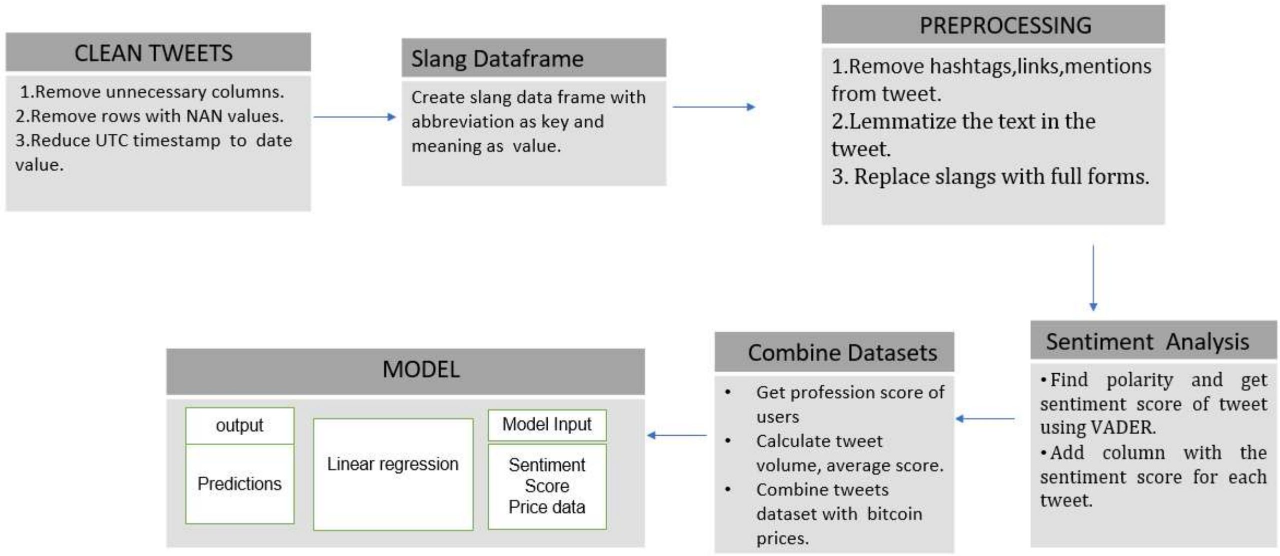


Fig 7.1-Architecture Diagram

7.2 Use case diagram

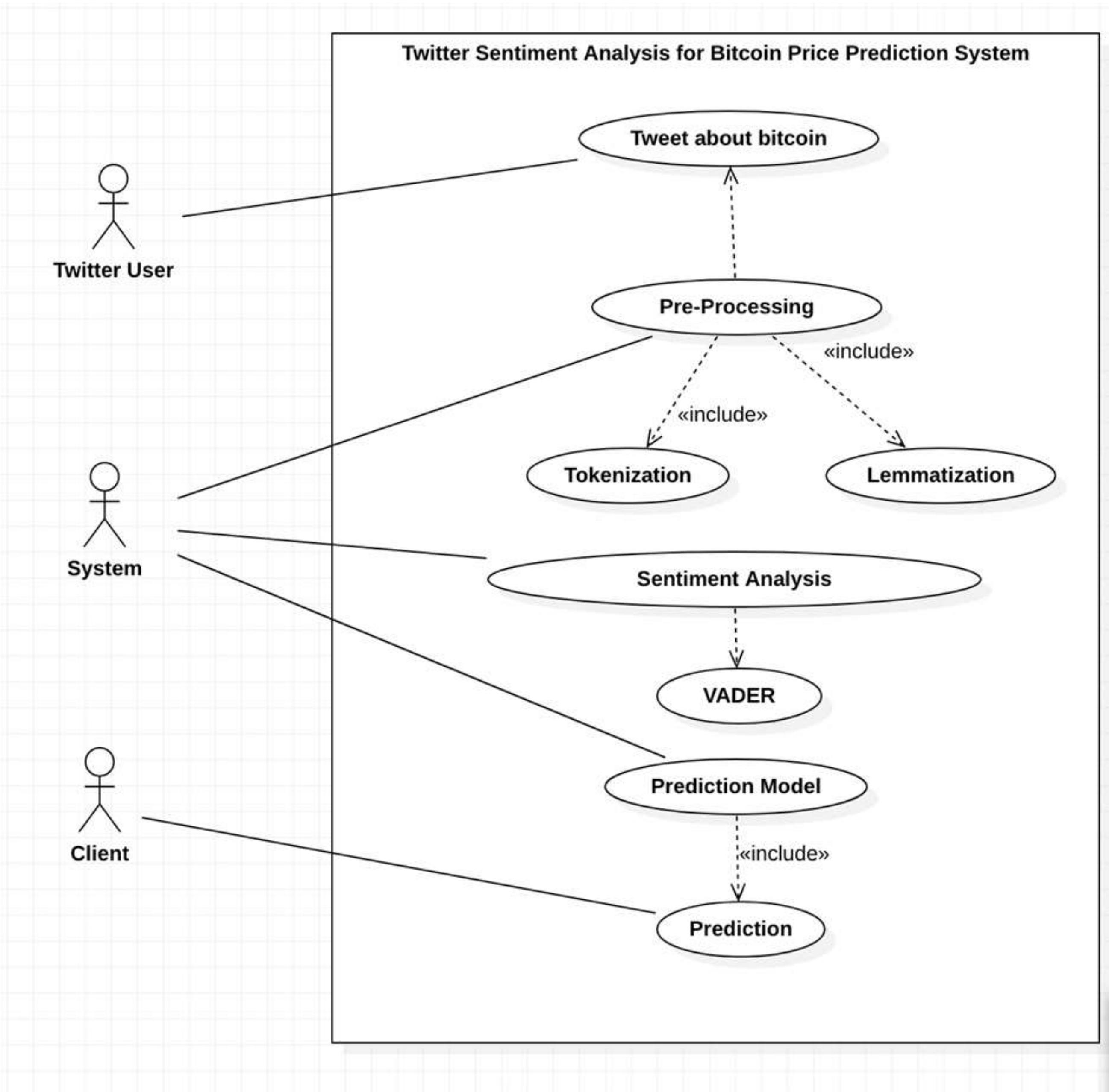
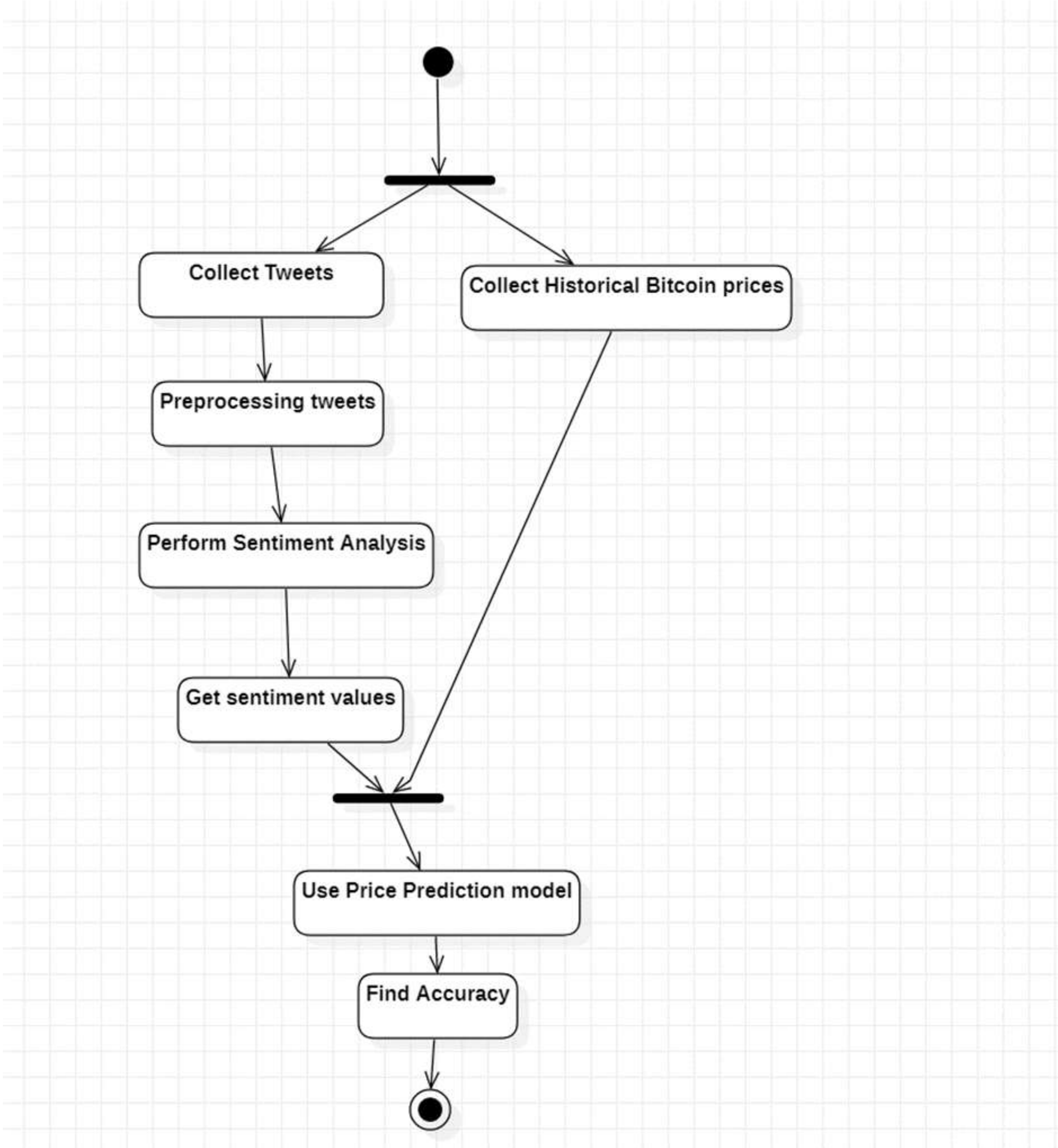


Fig 7.2-Use Case Diagram

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.

```
uncleaned_df=pd.read_csv("/content/gdrive/MyDrive/capstone project/datasets/tweets/tweets_2.csv")

uncleaned_df.drop_duplicates()
uncleaned_df=uncleaned_df.dropna(subset=['text'])
uncleaned_df=uncleaned_df[(uncleaned_df.user_verified==False) | (uncleaned_df.user_verified==True) \
                          | (uncleaned_df.user_verified=="True") | (uncleaned_df.user_verified=="False")]
uncleaned_df["user_verified"].replace({"True": True, "False": False}, inplace=True)

#drop is_retweet cause only "nan and false"
#drop hashtags,source,user_favourites,user_friends,"user_location",","user_created"
uncleaned_df.drop(["is_retweet","hashtags","source"],inplace=True,axis=1)
uncleaned_df.drop(["user_favourites","user_friends"],inplace=True,axis=1)
uncleaned_df.drop(["user_location","user_created"],inplace=True,axis=1)

#reducing UTC timestamp to only date
uncleaned_df["date"]=[i[:11] for i in uncleaned_df.date ]
```

Fig 8.1-Preprocessing the dataset

8.2 Create dataframe of slang data

- Slangs in the tweets must be expanded as 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.

```
def cleanText(string):
  try:
    fin=""
    string=string.replace('\n'," ")
    #print(string)
    string=string.split()
    for  i in range(len(string)):
      if (string[i]=="") or (string[i][0] in "@/") or ("http" in string[i]) or \
      (string[i][0] =="#" and string[i][-1] not in "?!.,"):
        continue
      else:
        if((string[i][0] =="#" and string[i][-1] in "?!.,")):
          fin+=" "+string[i][1:]
        else:
          fin+=" "+string[i]
    return fin.strip()
  except Exception as e:
    print("ERROR"," ".join(string))
    print(e)

def convertSlang(string):
  string=string.split()
  for i in range(len(string)):
    s = re.sub(r'[^\w\s]','',string[i])
    if(findInDict(s.lower())==1):
      string[i]=string[i].replace(s,SLANG_DICT[s.lower()])

  return " ".join(string)
```

Fig 8.2-Functions to process tweets

```
cleaned_text=[cleanText(i) for i in df.text]
slang_expand=[convertSlang(i) for i in cleaned_text]
lemmatized_text = [lem.lemmatize(t) for t in slang_expand]
df.text=slang_expand
#
sentiment=[sentiment.polarity_scores(convertSlang(i))["compound"] for i in lemmatized_text]
```

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 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.

```
"""
2 kinds of professions possible:
SINGLE- example Hassan Sajwani: BUSINESSMAN, present in class_name "g.wF4fFd.JnwWd.g-blk"
MULTI- example Ted Cruz:['POLITICIAN,SOLICITOR,CRITIC'], present in class_name "KKHQ8c"
returns peofession or "-"
"""
def singleProf(url):
  try:
    wd.get(url)
    ans=0
    ans = WebDriverWait(wd, 3).until(EC.presence_of_element_located((By.CLASS_NAME, "g.wF4fFd.JnwWd.g-blk")))
    if("Related search" in ans.text):
      return 0
    return ans.find_element(By.CLASS_NAME,"Z0LcW.t2b5Cf").text
  except:
    if ans:
      try:
        return ans.find_element(By.TAG_NAME, 'b').text
      except:
        try:
          return ans.find_element(By.CLASS_NAME,"ILfuVd").text
        except:
          return ans.text
    return 0
def multipleProf(url):
  try:
    wd.get(url)
    ans = WebDriverWait(wd, 3).until(EC.presence_of_element_located((By.CLASS_NAME, "KKHQ8c")))
    return ans.text.replace("\n",",")
  except:
    return 0

def getProfession(name):
  wd.refresh()
  name=deEmojify(unidecode(name))
  url=str("https://www.google.com/search?q="+"+".join(name.split(" "))+"+profession")
  #single prof
  result=singleProf(url) or multipleProf(url)
  if result:
    return result
  return "-"
```

Fig 8.4-Code to obtain the profession of the user

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

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.

```
def get_data(symbol, interval, start, end):
    data = yf.download(tickers=symbol, start=start, end=end, interval=interval)
    return data

symbol = "BTC-USD"
interval = "1d"
start = "2021-02-05"
end = "2022-04-18"
data = get_data(symbol, interval, start, end)

[*********************100%***********************]  1 of 1 completed
```

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 tweet data using the date as the key.

```
#tweet_data is the modified dataframe containing all the tweets
scores = []
for i, s in tqdm(tweet_data.iterrows(), total=tweet_data.shape[0],position=0, leave=True):
    try:
        scores.append(s["score"] * ((s["user_followers"])) *(s["profession_score"]+1))
    except:
        scores.append(np.nan)
tweet_data["final_score"] = scores
tweet_data.head()
```

Fig 8.7-Calulating tweet score

```
#btc_data => bitcoin prices
#tweet_aggregates => daily bitcoin tweet aggregates
combined_data=pd.merge(btc_data,tweet_aggregates,how="left",on="Date",sort=True)
```

```
combined_data.head()
```

	Date	Open	High	Low	Close	Tweet_volume	Avg_score
0	2021-02-05	36931.546875	38225.906250	36658.761719	38144.308594	1694.0	592.361387
1	2021-02-06	38138.386719	40846.546875	38138.386719	39266.011719	3278.0	498.539178
2	2021-02-07	39250.191406	39621.835938	37446.152344	38903.441406	3030.0	219.357551
3	2021-02-08	38886.828125	46203.929688	38076.324219	46196.464844	5647.0	1328.579672
4	2021-02-09	46184.992188	48003.722656	45166.960938	46481.105469	4350.0	1394.978194

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 target are standardized as they have varying ranges.
- Predictions obtained using Linear Regression and graph plotted.
- Model is validated using R^2-score and RMSE .

```
features = ['Open', "High","Low","Tweet_volume","Avg_score"]
target = ['Close']

#features & target for the model
featureScaler = StandardScaler()
targetScaler = StandardScaler()

featureScaler.fit(train_df[features])
targetScaler.fit(train_df[target])
print(featureScaler.mean_)
print(targetScaler.mean_)


(X_train, Y_train), (X_test, Y_test) = create_dataset(train_df, test_df)

# Logistic Regression baseline model
regressor = LinearRegression()

# for logistic regression, we will have to flatten the time dimension so that the data will be 2D (N, 10*7)
# so the model will have 70 inputs and 1 output
#train
regressor.fit(X_train,Y_train)

#test
predictions = regressor.predict(X_test)
loss = mse(predictions,Y_test)
print(f'MSE for Logistic Regression Baseline : {loss}')

pickle.dump(regressor, open('regressor.pkl', 'wb'))
pickle.dump(featureScaler, open('featureScaler.pkl', 'wb'))
pickle.dump(targetScaler, open('targetScaler.pkl', 'wb'))

compare_pred(predictions)
plot_predictions(predictions)
```

Fig 8.9- Model

8.8 Accuracy calculation

Training and testing accuracy of model is calculated

```
#training accuracy
regressor_predictions = regressor.predict(X_train)
regressor_r2 = calculate_r2(regressor_predictions, train_df['Close'])
print(f'R2 for Linear Regression model -> {regressor_r2:.2f}%')
b.append(f'{regressor_r2:.2f}%')
```

```
R2 for Linear Regression model -> 99.31%
```

```
# calculate R2 metrics

regressor_predictions = regressor.predict(X_test)
regressor_r2 = calculate_r2(regressor_predictions, test_df['Close'])
print(f'R2 for Linear Regression model -> {regressor_r2:.2f}%')
b.append(f'{regressor_r2:.2f}%')
```

```
R2 for Linear Regression model -> 97.75%
```

Fig 8.10- Accuracy calculation

CHAPTER 9

RESULTS AND DISCUSSION

The mean square error for linear regression is found to be 0.003896.

The training accuracy obtained is 99.31%.

The testing accuracy of predictions obtained for the linear regression prediction model is found to be 97.75%.

The results obtained are at par with initial estimates.

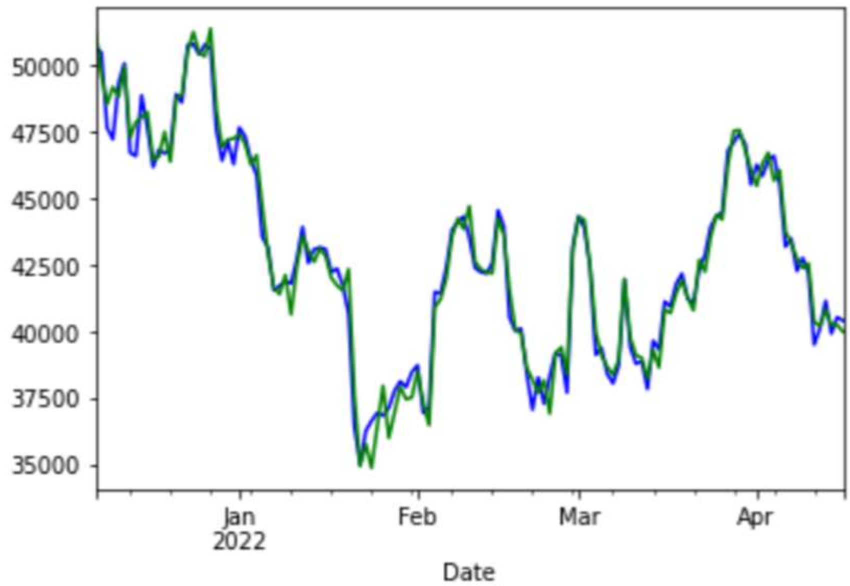


Fig 9.1- Predictions for price of bitcoin

In Fig 9.1, Green indicates the predictions for closing price while Blue is the actual closing price.

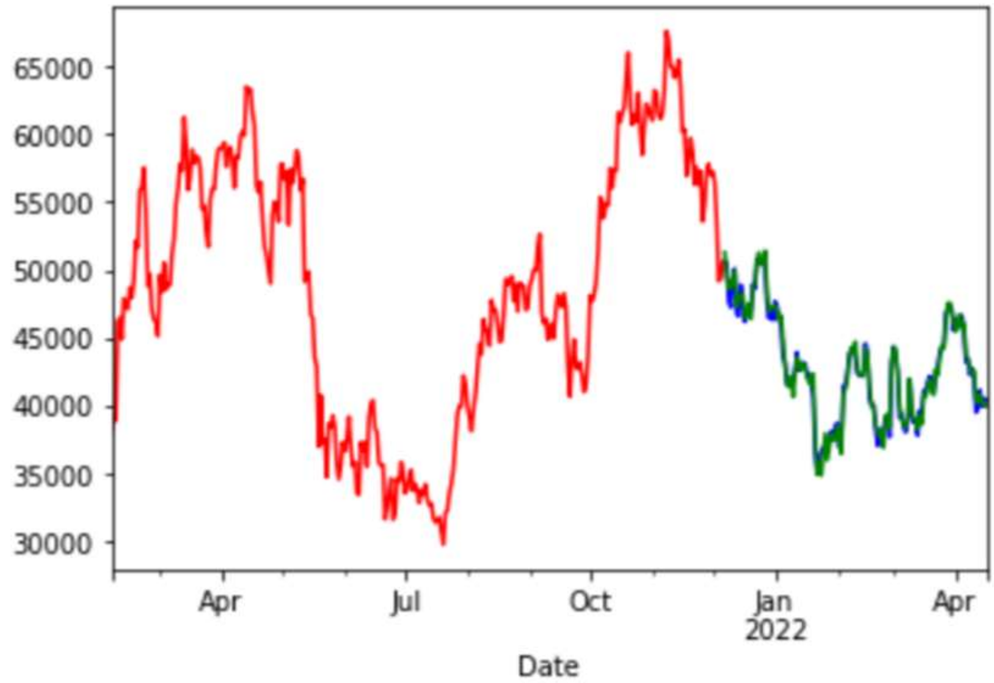


Fig 9.2- Bitcoin price of train and test data

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

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%

Fig 9.3- Metrics

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]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]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.

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

[4]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.

[5]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.

[6]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

[7]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

[8]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

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

[10]https://in.investing.com/crypto/bitcoin/historical-data.4

[11]https://online.visual-paradigm.com/

[12]https://keep.lib.asu.edu/items/133184

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

[14]https://helda.helsinki.fi/bitstream/handle/10138/344119/keningi_eino_thesis_2022.pdf

[15]https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

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

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 neural network class used to model sequential data.

VADER – “Valence Aware Dictionary and Sentiment Reasoner”, a sentiment analysing 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	Kaushal Mahajan	Namita Aluvithangal	Vedanth Mohan
Btech Computer Science	Btech Computer Science	Btech Computer Science	Btech Computer Science
PES University	PES University	PES University	PES University
Bangalore	Bangalore	Bangalore	Bangalore
achyut.jagini@gmail.com	kaushal.pes@gmail.com	namita0201@gmail.com	vedanth.pes@gmail.com

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.

II. DATA PREPROCESSING

The dataset used has 8 columns user_name,user_location,user_description,user_created,user_followers,user_friends,user_favourites,,user_verified, date, text, hashtags, source and is_retweet.

During cleaning "is_retweet", "hashtags", "source", "user_favourites", "user_friends" columns are removed.

Rows with NAN values are removed. Rows where "user_verified" value is neither True or False are deleted.

UTC timestamp is reduced to only date value.

Slang words in the tweets are expanded, hashtags ,mentions and links are removed from from the tweet .Lemmatization is applied to the tweet text.

Sentiment polarity scores are obtained for text for not lemmatized without slang, not lemmatized with slang ,lemmatized without slang and lemmatized with slang using VADER.

Polarity scores of ,lemmatized tweets with slang expanded is used as the sentiment score.

Using BeautifulSoup python module ,professions of verified users are added as a column in dataset

Using the users profession and a list of bitcoin related professions, a profession score is generated for each user. Each users' profession score is added as a column to dataset.

Final score of the tweet is obtained using sentiment score, user followers and profession score.

$$\text{Tweet\_ Score} = (\text{tweet\_sentiment\_score})*(\text{user\_followers})*(\text{profession\_score}+1) \quad (1)$$

Bitcoin price data is obtained for same period of the tweets from February 2021 to April 2022.

The tweet volume and average score of the tweets on each day is found .The values are normalized and combined with the bitcoin price data to get a combined dataset.

TABLE I.Tweet aggregates

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

III . LINEAR REGRESSION

Data split into training and test dataframe 70% -train.
Input features of model used are 'Open', 'High', 'Low', 'Volume', 'Tweet_volume', and 'Avg_score'
Target=['Close'].
The features and target 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).

IV.RESULTS AND DISCUSSION

The mean square error for linear regression is found to be 0.003896.The training accuracy obtained is 99.31%.The testing accuracy of predictions obtained for the linear regression prediction model is found to be 97.75%. The results obtained are at par with initial estimates.

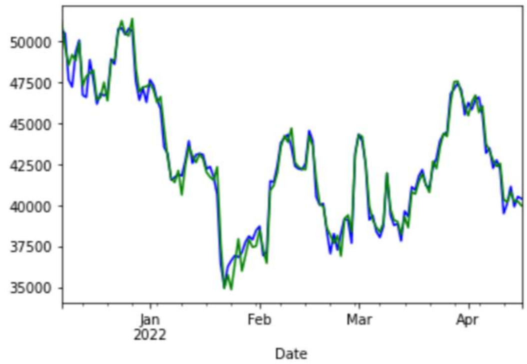


Fig 1. Predictions for price of bitcoin

In Fig 1, Green indicates the predictions for closing price while Blue is the actual closing price.

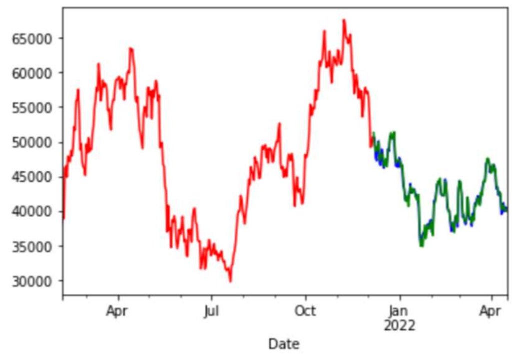


Fig 2- Bitcoin price of train and test data

In Fig 2, Red indicates the bitcoin price of training data and blue indicates bitcoin price of test data.Green indicates the predictions.

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%

Fig 3- Accuracy and metrics

V.CONCLUSIONS AND FUTURE WORK

The goal of the project 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 were obtained at par with initial estimayes

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

[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. *Financ Innov* 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.

ANNEXURE -II POSTER

Group No:55

Title: Twitter Sentiment Analysis for Bitcoin Price Prediction

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 analyse 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:

Achyut Jagini
PES2UG19CS013

Kaushal Mahajan
PES2UG19CS178

Namita Aluvathingal
PES2UG19CS245

Vedanth Mohan
PES2UG19CS449

Supervisor Dr Prajwala TR

Architecture/flow diagram