AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

SI 618: DATA MANIPULATION & ANALYSIS

Data Gathering & Processing Project Report - 1

Fall 2022

MOTIVATION

The Bitcoin cryptocurrency is a virtual currency that is designed to act as money and as a means of payment without being controlled by anyone, group, or entity, so that financial transactions are not reliant on third parties. There is a public ledger of every Bitcoin transaction, which makes it difficult to reverse transactions and impossible to fake them. Since Bitcoins are decentralized, they are not backed by a government or issuing institution, and their value is only guaranteed by the proof embedded within. The reason why it is worth money is simply that we, as people, decided it has value—same as gold. Bitcoin's value has risen dramatically since its launch in 2009. As of October 6, 1 bitcoin is worth approximately $20,019, despite once selling for less than $150 per coin.

There are many factors that influence Bitcoin's price, including supply and demand, investor and user sentiment, government regulations, and media hype. As a result of all these factors, price volatility is created. This project examines how different social media platforms, such as twitter and reddit, have responded in the past to these fluctuations in Bitcoin's price. Moreover, the objective of the project is not to establish any causal relationships among the tasks or any of their components, but it aims simply to draw hypotheses based on the observations that were collected as a result of the project.

THE DATASETS

At the outset of the project, the project proposal mentioned three datasets; however, upon further exploration of the datasets, it was found that there were some inconsistencies within the datasets, and after deliberation, the number of datasets was increased to four, as a result, the four datasets used in the project are:

1. **Bitcoin Tweets from 2021:** Tweets from 2009 to 2022 make up the first Twitter dataset. It is noteworthy that there are not many tweets between 2009 and 2020. Consequently, the tweets from 2021 through 2022 are considered in this analysis. The data totally consists of approximately 0.6M records with 13 columns, namely, user_name, user_location, user_description, user_created, user_followers, user_favorites, user_verified, date, text, hashtags, source and is_retweet.

   URL: Bitcoin Tweets | Kaggle

2. Bitcoin Tweets 2016 - 2019: To compensate for the pitfalls present in the first twitter dataset, the second dataset was taken. From the beginning of the year 2016 until the beginning of 2019, tweets were collected and analyzed in this dataset. This dataset consists of tweets scraped from twitter that contained keywords such as Bitcoin or BTC, collected using tools like Tweepy and Twint. There are approximately 16 million tweets in this database. Each tweet contains a user, fullname, tweet-id, timestamp, URL, likes, replies, retweets, as well as text.

   URL: Bitcoin tweets - 16M tweets | Kaggle

3. Reddit Comments: 4M+ Comments from Reddit that contain the word "bitcoin" from 2009 to 2019 collected from Google BigQuery. This dataset consists of 9 columns, namely, datetime, date, author, subreddit, created_utc, score, controversiality and comment text.

   URL: Reddit Comments Containing "Bitcoin" 2009 to 2019 | Kaggle

4. Bitcoin Historical Prices: The data in this dataset was taken from Yahoo Finance, which contains historical Bitcoin prices. Bitcoin prices from 2015 to 2022 are included in this dataset. There are fields such as date, open, high, low, close, adj.close, and volume included.

   URL: Bitcoin USD (BTC-USD) Price History & Historical Data - Yahoo Finance

MANIPULATION AND JOINS

The datasets used in the project were all in .csv format. In order to analyze the datasets, they were simply downloaded from their respective sources and uploaded on the GreatLakes Server, where they were imported into the project's Jupyter notebook using SparkSQL.

Consequently, after the datasets were imported into the project notebook, the schema of each dataset was checked, to check for the data types. It was noticed Spark considered all the attributes as String data type by default. Hence, the following data types were changed using *.within()* function, from respective tables,

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

Prices

Attribute	Modified Data Type
Date	DateType()
Open	DoubleType()
High	DoubleType()
Low	DoubleType()
Close	DoubleType()
Adj. Close	DoubleType()
Volume	DoubleType()

Tweets 2021

Attribute	Modified Data Type
Date	DateType()
user_followers	IntegerType()
user_favorites	IntegerType()
user_friends	IntegerType()
user_verified	BooleanType()
is_retweet	BooleanType()

Reddit

Attribute	Modified Data Type
Date	DateType()
score	IntegerType()
controversiality	IntegerType()

Tweets 2016

Attribute	Modified Data Type
Date	DateType()
replies	IntegerType()
like	IntegerType()
retweets	IntegerType()

Fig. 1: Changing Data Types of Attributes

Furthermore, the Twitter 2021 dataset contained text from the different tweets as well. Since, the body of the tweets can contain special symbols, emojis, user mentions, links etc., it becomes crucial for us to clean the tweet texts. Therefore, multiple forms of regular expressions (RegEx) were used to remove such discrepancies in the text attribute.

After changing the datatypes, and cleaning the datasets as per the requirement for the computational tasks, the datasets were saved as a Spark Temporary Tables using the *.registerTempTable()* function.

A date and time stamp are included in each dataset, which includes tweets, reddit comments, bitcoin prices and the number of bitcoin wallets. The tweets come from two different datasets, derived from two different sources. Though the time frame of both of these datasets is different, both contain a date attribute. The datasets all cover a roughly similar date range between 2016 and 2022. As a result, these datasets were be merged according to the month and year they were collected. Spark SQL was used to combine these datasets using inner joins as and when required during the computational tasks.

Furthermore, there were some rows where the dates extended beyond the 21st Century. The rows containing such dates were simply excluded from the analyses, only the dates in the interval of years 2016 and 2022 were considered. Moreover, amongst other discrepancies, there were some instances where the text of tweets were blank, such tweets were also removed from our analyses.

Moreover, some attributes had values larger than the others, this needed some standardization in order to visualize the dataset better. Therefore, MinMax Scaler was used to standardize such values between 0 and 1. The scaling in turn made the interpretation of graphs a bit easier.

The libraries used in the project were: pyspark.sql.functions, sklearn.preprocessing, pandas, nltk, stopwords, wordcloud, matplotlib, collections.

ANALYSIS AND VISUALISATIONS

As described in the project proposal, there are four computational tasks that were the main focus of the project. All the computational tasks focus on different aspects of the datasets and are inclined towards using different combinations of datasets to derive relevant insights.

Computational Task 1: How do fluctuations in the prices of Bitcoin affect the volume of tweets and comments across different platforms?

In order to understand how and when Bitcoin began to become the talk of the crypto market, it might be useful to examine the volume of tweets and comments generated after price fluctuations, which indicates its rise and maybe, fall in popularity as an increasing number of different currencies are being offered to investors on the crypto-market.

The main focus of this task is to observe how different social media platforms, in this case twitter and reddit are affected, how the number of tweets and reddit comments increase or decrease with the fluctuations in the price of Bitcoin between the years 2016 and 2019? To accomplish this task, the datasets that were used in this task were: prices, tweets2016 and reddit.

The steps involved in the task included:

i. Joining the three datasets to get a data frame that contained date, count of tweets, count of reddit comments and price of bitcoin. This data frame was grouped by the date attribute to get a total count of tweets and comments for a particular date.
ii. The Spark Dataframe was converted to Pandas DataFrame. The counts of tweets & comments and price were then standardized using MinMax Scaler.
iii. Line Plot was then used to plot all the different attributes on a single graph.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

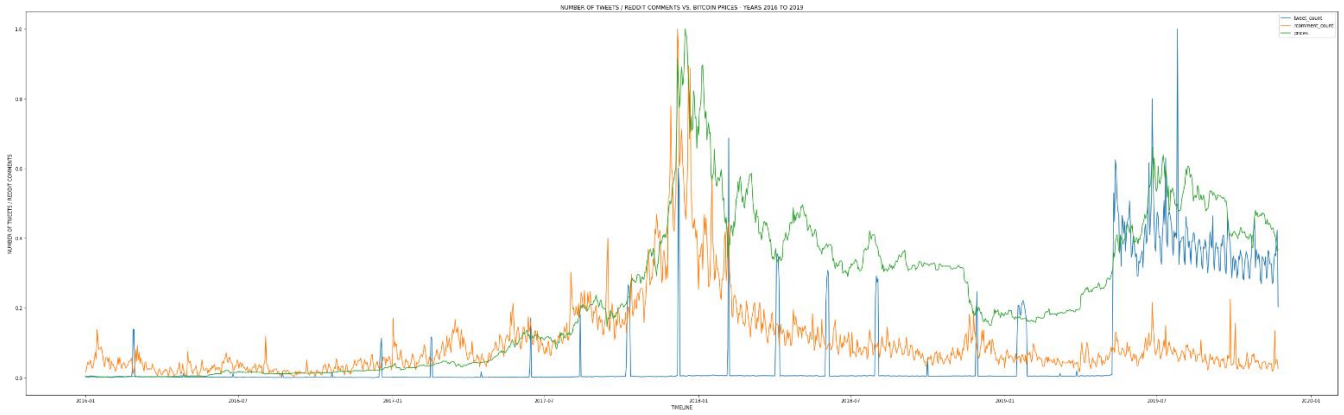

Fig. 2: Number of Tweets / Reddit Comments VS. Bitcoin Prices - Years 2016 to 2019
(For detailed view, refer fig2.png attached with the report)

Fig. 3: Legends for Fig. 2

From the above visualization, we can observe that we have different trends for reddit comments and the number of tweets. The reddit comments show a trend significant with the price fluctuations. However, the number of tweets show seasonal spikes until mid-2019, this could either be due to the inconsistency present in the collection of tweets in the dataset. However, from the reddit comments, we can observe that the number of comments does increase with fluctuations in the prices of Bitcoin.

Moreover, after July 2019, we see a significant trend in the number of tweets. It could be indicative of the fact that the with high fluctuations in the prices of Bitcoin, the number of tweets does tend to increase. We also observe that following the early 2018, there is a decrease in the number of reddit comments that are related to Bitcoin.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

Computational Task 2: What are the most common keywords that are used in the tweets? Can some words be identified could have caused a positive or a negative impact?

Over the past decade, Bitcoin has been in the news and all-over social media since its inception in 2009. It has gone through good days and bad days. The analysis of some common positive and negative keywords from tweets and comments on reddit could provide insight into how different words developed over time. For this analysis, the main years of focus are the years 2021 and 2022. The datasets used for this task are: prices and tweets2021.

The main challenges faced in this task were to clean the text of the tweets, which involved creating and using multiple regular expressions. Moreover, there were some domain specific keywords that needed to be removed from the texts in order to focus on the relevant keywords needed for the analysis of the text.

The steps involved in the task included:

i. Creating a list of stopwords, using the nltk library. Stop words are a group of frequently used terms that are used in all languages, not just English. Stop words are essential in many applications because they allow us to concentrate on the relevant words rather than the words that are overused in a language.
ii. Amongst the already specified stop words in the nltk library, some domain specific stop words were also added to the list. Furthermore, this list of stopwords was converted into a Spark Data Frame.
iii. Percentage change in the price of Bitcoin was calculated for each date using the opening price and the closing price. The top 10 percentage change in the prices were used for these analyses.
iv. The data frame was then joined with the tweets2021 dataset in order to get the tweets that were posted on these dates. The top 10 positive percentage change were used to analyze for positive keywords and similarly, the top 10 negative percentage change in the prices made up for analysis of the negative keywords.
v. The text of the tweets was segregated into different words using the *.explode()* function and then counted to get a total count of different keywords. Furthermore, the stop words were removed from this list of words in order to focus on relevant keywords.
vi. A Word Cloud was formed for positive and negative keywords, which represents the frequency with which words occur in the tweets.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

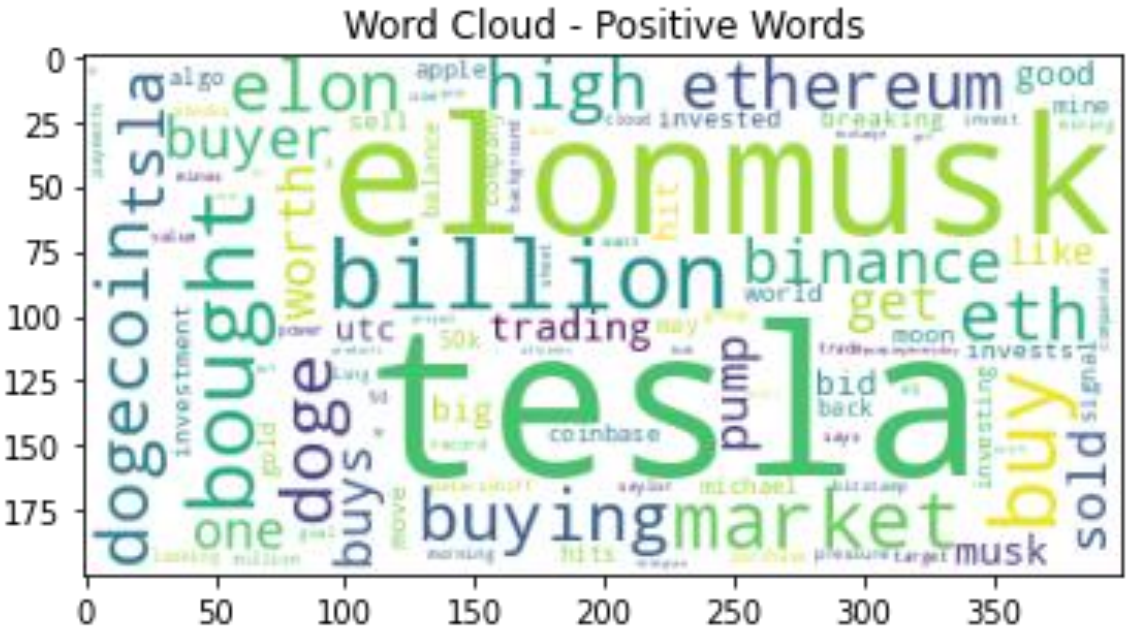

Fig. 4: Positive Keywords from 2021 & 2022 Tweets

From the above visualization of positive keywords, we can see that some of the prominent words that occurred in the tweets were: elonmusk, tesla, dodgecoin, billion, binance, ethereum and many combinations of the word “buy”.

These keywords could be explained by the fact that during the early 2021, Tesla's initial acquisition of $1.5 billion worth of bitcoin was big news, and the price of BTC increased by more than 60% in the month that followed the announcement. During that time the sentiment of the market was high, which encouraged strong buying of the cryptocurrency.

The value of the cryptocurrency has fallen precipitously in 2022, nevertheless, following a significant gain in 2021. As the value of the currency increased, the corporation swiftly sold 10%, generating a profit of $101 million during the period. Due to the negative effects of mining on the environment, Musk stated in May of last year that Tesla will stop taking bitcoin for car purchases, but he also stated that the business "will not be selling any bitcoin."

Relevant news articles:

- Does Crypto Still Care About Elon Musk? (coindesk.com)
- Tesla buys $1.5 billion in bitcoin, plans to accept it as payment (cnbc.com)

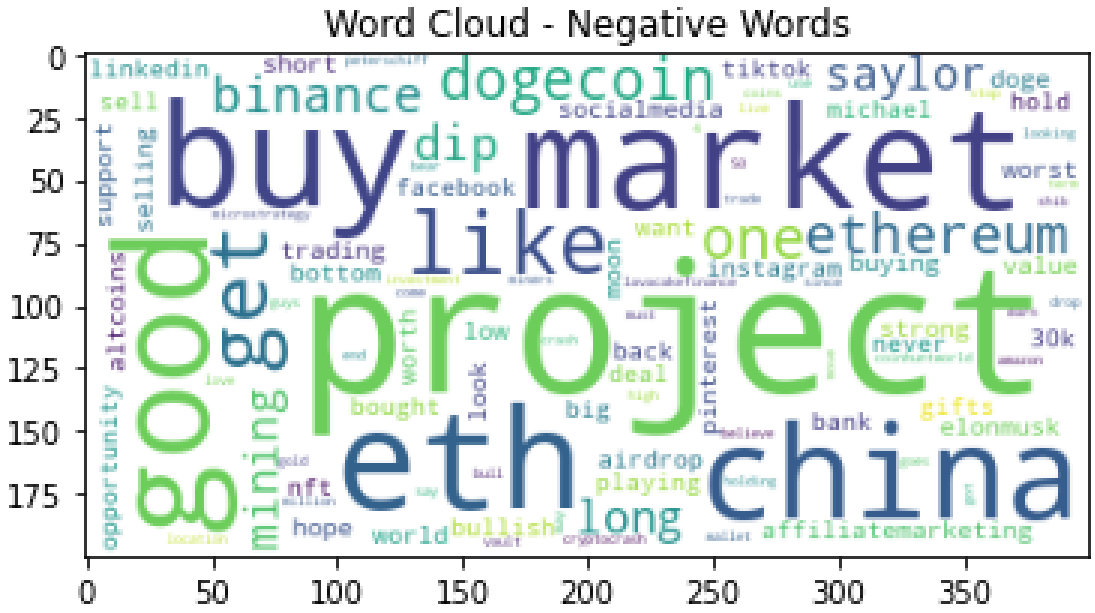

Fig. 5: Negative Keywords from 2021 & 2022 Tweets

From the above visualization of negative keywords, we observe that some of the prominent words that were noticed in the tweets were: project, China, market, Ethereum, airdrop, Michael Saylor etc.

The occurrence of the word China may be related to the People's Bank of China (PBOC) decision to outlaw all cryptocurrency transactions in late September 2021. Due to their highly speculative nature, cryptocurrencies have been linked to financial crime and are posing an increasing risk to China's financial system, according to the PBOC.

In addition, Michael Saylor's Bitcoin holdings had lost almost $1.3 billion in value due to the Bitcoin meltdown in early September 2022, which led to a tax fraud lawsuit against him.

Relevant news articles:

- After Chinese Ban, Cryptocurrency Mining Got Worse for Climate - The New York Times (nytimes.com)
- The Bitcoin crash has wiped out over $1.3 billion in value from Michael Saylor's Bitcoin holdings. Now he's being sued for tax fraud | Fortune

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
October 27, 2022

Computational Task 3: How often do people with verified twitter accounts tweet regarding the fluctuations in the prices of Bitcoin?

In particular, it would be interesting to see how influential people respond to different changes in the Bitcoin market price based on their reach on social platforms. There might be some interest in discovering when did people with verified accounts start to talk about Bitcoin since such people tend to be more deliberate with what they post than the general audience. This would be helpful in getting a more realistic view, which should, hopefully, not be influenced by fake news, and hypes.

Conversely, it could very well be the case that tweet from a verified user leads to the fluctuation in the price of Bitcoin, but it is not the focus of this task. For this task, we mainly focus on the post-fluctuation analysis. The datasets used for this analysis are: prices and tweets2021.

The steps involved in this task included:

i. Counting the number of tweets made by verified people and on which dates.
ii. The data was then merged with the closing prices of Bitcoin with respect to the dates.
iii. The Spark Data Frame was then converted to Pandas Data Frame and then MinMax Scaler was used to standardize the values.
iv. Line Plot was used to visualize the number of tweets made by verified users along with the fluctuations in the price of Bitcoin
v. To further investigate the spike in the number of tweets, bar graph was used to visualize the number of tweets made on particular dates.

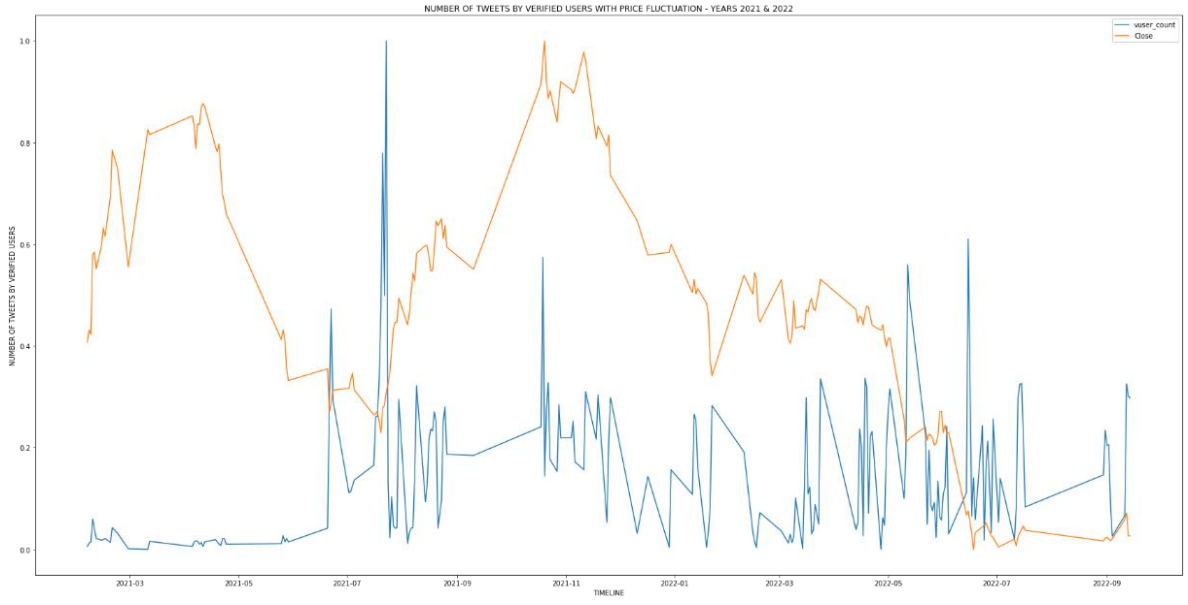

Fig. 6: Number of Tweets by Verified Users with Price Fluctuation
(For detailed view, refer fig6.png attached with the report)

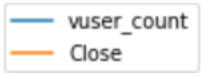

Fig. 7: Legends for Fig. 6

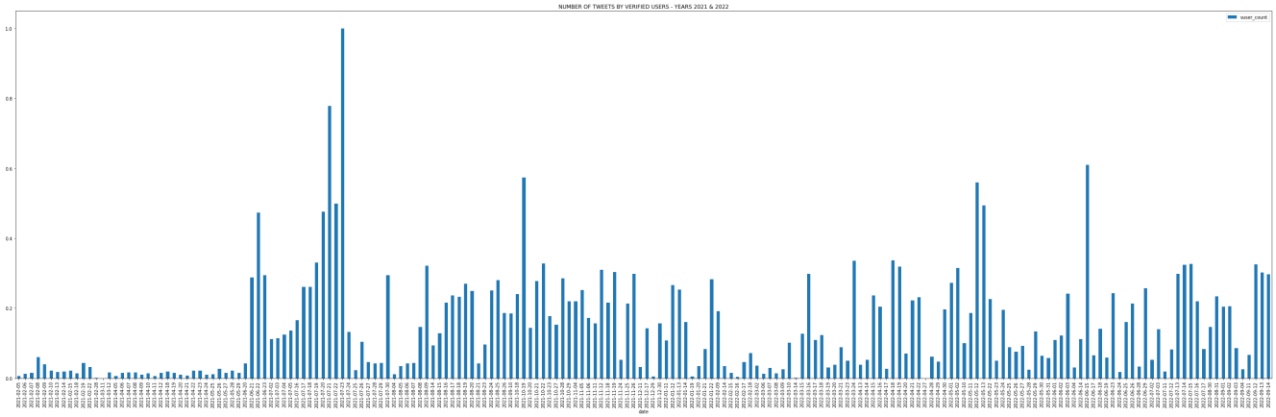

Fig. 8: Number of Tweets by Verified Users in the Years 2021 & 2022
(For detailed view, refer fig8.png attached with the report)

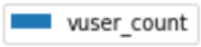

Fig. 9: Legend for Fig. 8

From the line graph, we can hypothesize that verified users do tend to be selective with their tweets. For instance, we can see that there is a sharp decline in the price of Bitcoin late in the month of April 2021, but there is no significant amount of spike in the number of tweets. However, a little fluctuation in the price in late June 2021, the number of tweets suddenly sees an increase. We see a similar trend in the later during the end of the month July in 2021. To further narrow down specific dates, we look at the bar graph which gives us a more precise idea about the number of tweets. From the bar graph we can see that a significant number of tweets were made on the dates 2021-07-23, 2021-07-21, 2021-10-19 & 2022-06-15.

CHALLENGES ENCOUNTERED

One of the major challenges encountered was that the inconsistency present in the twitter dataset. The initial twitter dataset that was proposed in the project proposal, had all the columns needed for the analysis. However, it turned out that the number of tweets present in the data for

the years 2013 to 2020 were very less (less than 100 for some years). This made the analysis extremely difficult with such less tweets.

The inconsistency with the data was countered by adding another data which contained the tweets from the years 2016 till 2019. However, the second dataset did not contain the tweet text. Moreover, even after taking the second dataset into the analysis, the tweets for the year 2020 are still missing. It is one of the main reasons that the analysis is not consistent when it comes to the timeline. Therefore, some of the computational tasks needed to be changed so as to fit the datasets.

The second challenge encountered was with the stop words. Using nltk library, though we were able to eliminate some common words from the Word Cloud analyses, to remove some domain specific words, brute force was used to identify stop words manually. Even after much deliberation and research, stop words related to the domain of cryptocurrencies or bitcoin were not found.

Lastly, since the twitter datasets are inconsistent, the data dictionary does not mention how exactly was this data scraped off, which limits our analysis, making us unable to make any causal inferences with the data.

REFERENCES

- Bitcoin Magazine - Bitcoin News, Articles and Expert Insights
- Prediction of Bitcoin prices using Twitter Data and Natural Language Processing.pdf (duke.edu)
- Spark SQL Built-in Standard Functions - Spark by {Examples} (sparkbyexamples.com)
- How to Use StandardScaler and MinMaxScaler Transforms in Python (machinelearningmastery.com)
- 140+ Blockchain and Crypto Words: The Ultimate A-Z Glossary | FinTech Magazine
- Removing stop words with NLTK library in Python | by Banjoko Judah | Analytics Vidhya | Medium
- PySpark Word Count. Apache Spark is an open-source… | by Gülcan Öğündür | Medium