Sentilyzer : Lexicon based Sentiment Analysis of Tweets

Monica Verma, Gunjan Mishra, Khushboo Kumari, Inteyash Ahmed, Kshama Mishra

Department of Computer Science and Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh, India

***Abstract*— Sentiment analysis is a process of deriving emotions from texts. Twitter Sentiment Analysis is used to find market insights and trends or changes in people's perceptions contained in tweets mentioning particular product or service on Twitter which is a popular social media platform. As the world is slowly shifting towards digital currency, we have seen that social media is critical in establishing public perception about cryptocurrency and crypto market trends. The purpose of this study is to determine if there is any correlation between positive or negative perceptions of cryptocurrencies and their value, and whether its popularity has changed over time. We have used lexicon-based approach for analyzing the sentiments of tweets referring to either of the two most popular cryptocurrency namely Bitcoin and Dogecoin. We extracted tweets from Twitter from different geographical locations and different time periods in order to make our sample size as diverse as possible so that we could get precise results. For visualization of these results, we have created dashboard in tableau. These visualisations show popularity of BTC and Doge through total count of tweets, their polarity i.e. negative, positive or neutral, sentiment scores, etc. Our study has shown that though there are times when Dogecoin's value has significantly shot up due to positive tweets especially by most influential figures like Elon Musk but the effect isn't permanent and popularity of Bitcoin and Dogecoin remains volatile.**

***Keywords*— lexicon, tweets, opinion, analysis, map, time, location, emotion, sentiment, score, polarity, visualisation, tableau, etc.**

I. Introduction

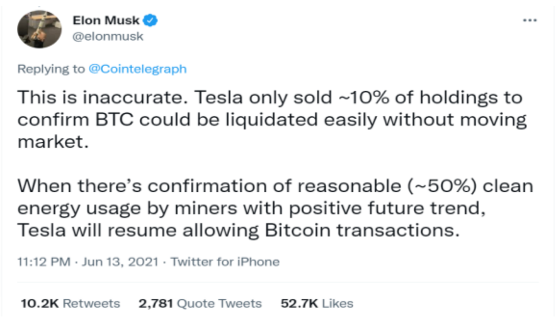
Twitter sentiment analysis provides means to analyse tweets in real-time, and determine the sentiment that underlies each message. In this research paper, we have decided to focus on validating whether there is any correlation between positive or negative sentiment towards two of the most popular Cryptocurrency namely, BTC and Doge and its price & popularity. Cryptocurrency is any form of currency that exists virtually and doesn't have a central issuing or regulating authority. It issues new units by analysing supply and demand. This demand is severely affected by media hype and people's sentiments and perception towards it which they tend to express online.

Here is the introduction to two of the most popular crypto coins and how people's sentiments over Twitter have affected their price and popularity in past:

A. Bitcoin (BTC):

Bitcoin was firstly introduced in 2009 and became popular around 2010. Since then, its price has been increasing and sometimes falling thousands of dollars within days and stands at $29,399.89 USD as of 14th May 2022. But the nature of cryptocurrency is volatile, i.e., its price fluctuates based on current supply and demand.

The most recent example of this volatile nature can be seen by Elon Musk's tweets in the past. Elon Musk is the most followed CEO on Twitter with 92 million followers and is the wealthiest person in the world as of April 2022 and hence a very influential figure for people online. In May, Musk made an announcement over Twitter that Tesla would no longer accept bitcoin as payment due to environmental concerns about its heavy energy use. As a result, the price of bitcoin plummeted to around 15% of its previous value. Before that tweet, the value of the whole cryptocurrency market stood at around $2.43 trillion, according to data from Coinmarketcap.com. Within a few hours of his tweet, the market capitalization had dropped to around $2.06 trillion, resulting in loss of around $365.85 billion. After few days, Musk again clarified the whole Bitcoin situation in a reply to a tweet (Fig 1) suggesting that Tesla would again accept bitcoin once miners moved to clean energy usage which brought the cryptocurrency's price up 8 percent.



Elon Musk
@elonmusk

Replying to @Cointelegraph

This is inaccurate. Tesla only sold ~10% of holdings to confirm BTC could be liquidated easily without moving market.

When there's confirmation of reasonable (~50%) clean energy usage by miners with positive future trend, Tesla will resume allowing Bitcoin transactions.

11:12 PM · Jun 13, 2021 · Twitter for iPhone

10.2K Retweets 2,781 Quote Tweets 52.7K Likes

Fig 1. Tweet by Elon Musk (Source: twitter.com)

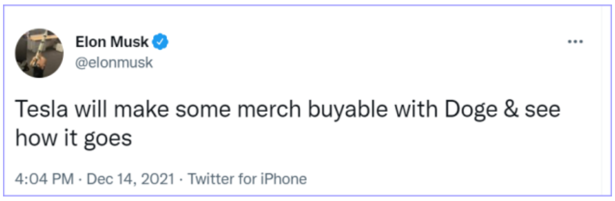
B. Dogecoin:

Dogecoin was initially started as a joke by its creators Billy Markus and Jackson Palmer in 2013. It's named after Japanese Shiba Inu dog, which inspired the original "doge" meme.

Elon Musk and some other people praised dogecoin and said that it has more potential to be an actual currency than bitcoin. Tweets from Musk and surging memes over Twitter and other social media platforms fueled interest in the cryptocurrency in 2021, causing some of his followers to buy this crypto and driving its price higher than ever before.

Dogecoin went from being worth just a fraction of a penny at the start to notching a record-high price above 74 cents in May, just after Musk announced on Twitter that he is working with developers to improve doge's system transaction efficiency.

But many times, negative perception about it in media has caused a lot of loss to its investors. E.g. In 2021, Elon Musk appeared on the American comedy show, Saturday Night Live, where he referred to Dogecoin, as a "hustle". Dogecoin has lost most of its value since, with the mass selling of the coin. Again, after tweet from Elon Musk in December 2021, in which he said that Tesla would accept Doge as payment [Fig. 2], the cryptocurrency skyrocketed as much as 17.36 U. S Cents but now has again declined to 8.43 U.S. cents as of May 2022.



Elon Musk
@elonmusk

Tesla will make some merch buyable with Doge & see how it goes

4:04 PM · Dec 14, 2021 · Twitter for iPhone

Fig 2. Tweet by Elon Musk (Source: twitter.com)

Hence, we can concur this meme-inspired cryptocurrency has very high price fluctuations mainly because of people's perception of taking this currency as a joke or as a serious future investment and in this Social media does play a significant role because many people rely on them to decide which crypto currency to invest in.

To reach a concrete conclusion, we have performed sentiment analysis on the tweets collected from December 2021 and May 2022 in order to see the correlation between people's sentiment towards BTC and Doge and to analyse if their popularity has changed over time or not.

II. Literature Survey

A. Literature Review

Chahat Tandon et al, (2021) categorised the tweets on cryptocurrency and predicted the price forecasting using ARIMA model. They succeeded in predicting 10 future values of bitcoin with 96% accuracy and 0.0395 average errors. [1]

Shathik et al, (2020) considered various techniques to do polarity analysis of data. Machine learning Basic sentiment analysis technique emerged as the most successful algorithm for this task. [2]

R. Liu et al., (2019) considered a new method of machine learning namely Transfer education that utilizes existing knowledge to solve problems and also used to make predictions from data but since certain aspects of sentimental analysis through Natural Language Processing has not been explored yet hence using transfer learning has proved to be difficult. Although we can conclude from this study that in future there is a promising scope for aspect level sentiment analysis.[3]

Gopu & Swarnalatha et al., (2017) performed comparative study of sentiment analysis of product reviews using various machine learning algorithms such as bag-of-words, n-gram, natural language processing and naive Bayes classifier. This study was primarily done to find out the polarity towards the product and which features have the most positive feedback in order to know which product's feature to focus on. In future, this study can be further extended by analysing opinions in different languages and text containing slangs and sarcasm which are hard to detect by current analysis technique. [4]

J. Singh et al., (2017) has focused on sentiment analysis of online media such as movies, products, and customer's opinion on online platforms and how it's affecting people. Since sentiment analysis in natural language processing is used to find positive and negative polarities from data collected from social media, these services can be used to help in determining clients of a particular product or movie etc. so that the future products can be catered to their needs. For this, we can collect the data from blogs, forums, tweets, articles, feedback forms etc. [5]

Mika Mantyla et al.,(2016) found out from his study how quickly sentiment analysis is evolving since online social media platforms have came into existence. With the rise in social media, online product reviews and feedback became readily available which can be used to perform sentiment analysis. In 10 years, online data such as articles has increased about 50 times. Hence, sentiment analysis is ever -growing topic and is proved to be one of the most important topics for research purposes. [6]

B. Limitations and conclusion:

The task of sentiment analysis, especially in the domain of social media, is still in the developing stage and far from complete. Most of the work on sentiment analysis is based on machine learning models and the disadvantage of machine learning models is that they depend on labelled data. It is much difficult to ensure that proper accurate labelled data is acquired. So we have decided to use lexicon based approach instead which may result in further improved performance. Lexicon-based approach can be better understood and manipulated by a human and also it is easier to generate an appropriate lexicon than collecting labelled data.

III. Proposed Work

A. Problem Statement

We want to analyse two of the most popular cryptocurrencies namely Bitcoin and Doge in order to investigate: -

❖Which cryptocurrency has more reliable Twitter data available?

❖Which cryptocurrency is more popular with Twitter users?

❖How does the sentiment look over the past few days?

❖Are the polarity patterns different among different cities?

- Popularity of BTC and Doge
- Popularity in different locations on a map (Bangalore, California and San Francisco)
- Changes in popularity over time (Comparison between December 2021 and May 2022 data)

❖Are Bitcoin and Doge volatile and affected by social media sentiments and number of tweets?

B. Methodology:

We have created a function to clean the tweets and then measure sentiment score from them using lexicons and then created a dashboard for visualisation using Tableau. It shows the number of tweets, the polarity of tweets, popularity of BTC and Doge on maps and the sentiment scores of tweets of in two different cities. Also, it displays the time span chart which shows change in their popularity over time.

Our approach of sentiment analysis involves following steps: -

1) Twitter Developer Account: One key ingredient for successful tweet harvesting is the Twitter developer's account. This one allows you to connect R with Twitter. It provides you with tokens and URLs to connect to the Twitter API.

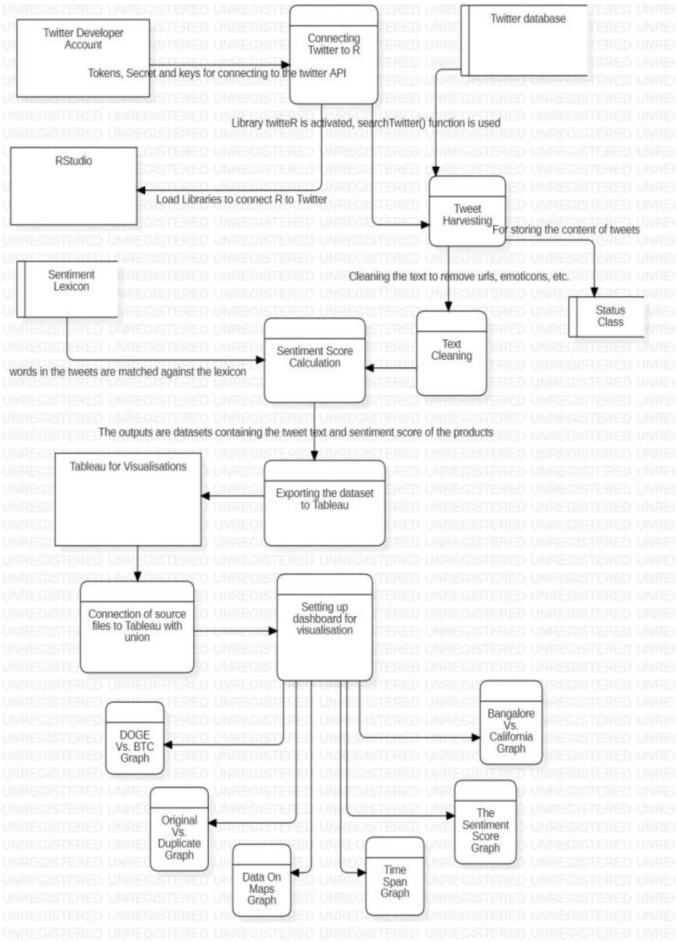

Twitter Developer Account
Connecting Twitter to R
Twitter database
Tokens, Secret and keys for connecting to the twitter API
Library twitteR is activated, searchTwitter() function is used
RStudio
Load Libraries to connect R to Twitter
Tweet Harvesting
For storing the content of tweets
Sentiment Lexicon
Cleaning the text to remove urls, emoticons, etc.
Status Class
Sentiment Score Calculation
Text Cleaning
words in the tweets are matched against the lexicon
The outputs are datasets containing the tweet text and sentiment score of the products
Tableau for Visualisations
Exporting the dataset to Tableau
Connection of source files to Tableau with union
Setting up dashboard for visualisation
DOGE Vs. BTC Graph
Bangalore Vs. California Graph
Original Vs. Duplicate Graph
The Sentiment Score Graph
Data On Maps Graph
Time Span Graph

Fig. 3 Data Flow Diagram for Twitter Sentiment Analysis and Visualisation

2) Connect R to Twitter: Interestingly, setting up a proper connection is one of the hardest things in the whole process. At least two systems communicate with each other which always leave a lot of space for problems. We need to provide a set of codes called Tokens, Secrets and keys which we got from twitter developer Account. This whole process is called a handshake.

These should be in proper format: -

1. Tokens, keys and secrets are string values hence they should be highlighted in green
2. Wrapped in quotes
3. No leading or trailing space

3) R Libraries: Install Packages to Connect R and Twitter and to process tweets:

A: Packages to Connect R and Twitter:

1. Library (RCurl1) (+ dependable packages):- General Network Client Interface for R

2. Library (httr) (+ dependable packages):- Contains functions to work with URLs and tokens

3. Library (twitter): - A collection of functions to harvest information from Twitter (e.g., tweets, user info, meta data)

B: Packages to Process Tweets: -

4. Library (tm): Text mining and natural language processing (e.g., cleaning and formatting text, changing format classes, text based filtering)

5. Library (stringr): Contains a tool to split tweets in single words

6. Library (dplyr): A tool for working with data frame like objects

4) Tweet Harvesting: First R is connected to the Twitter API, and then Library (twitteR) is downloaded and activated. After that we use Function searchTwitter().We can specify:

- ➢ Number of tweets to be harvested
- ➢ Search by Twitter user
- ➢ Language of tweets according to ISO-639 –I LANGUAGE NORM.
- ➢ Arguments for specifying a time span - It cannot exceed Twitters 14 days limit
- ➢ Specifies a geographical location i.e tweet should be from latitude, longitude and radius (km or mi)
- ➢ ResultType allows you to harvest either the most recent tweets or the most popular ones "popular", "recent".

5) Text Cleaning: This step is mandatory because: -

- Cleaned text can be processed by analytical tools (R, Tableau).
- Faulty text likely corrupts the analysis and calculations.
- URLs and emoticons can cause errors.
- Here we use gsub() function for cleaning the text.

6) Calculating Sentiment Scores: A major ingredient for our lexicon-based sentiment analysis is the opinion lexicon. This is simply a collection of words which the text is compared against. Positive lexicons include words like upscale, useful, great, etc. while negative lexicons include worst, worry, worthless, etc. Lexicons are language specific. In this Sentiment Analysis is done in English Language.

Two lexicon approaches:

1. positive lexicon hit gives a +1 score
2. negative lexicon hit gives a -1 score

Hits are counted and summed up to get a sentiment score. The libraries used are stringr and plyr. The tweets gets splitted into single words with str_split() function. A data frame named sentiment.df is obtained from previous calculations. This dataframe contains the tweet text and the calculated sentiment scores.

7) Exporting the dataset: The dataset is exported as .csv file. Tweettext, tweetdate, isretweet, retweetcount, favourite count, scores, product, city, country are included in the dataset. The duplicate tweets as well are present. So, we mark them with Boolean values

True=Duplicate

False=Original

The Duplicate column is added to the dataset. A .csv file is created for every city and each category (E.g., Doge Coin and Bitcoin)

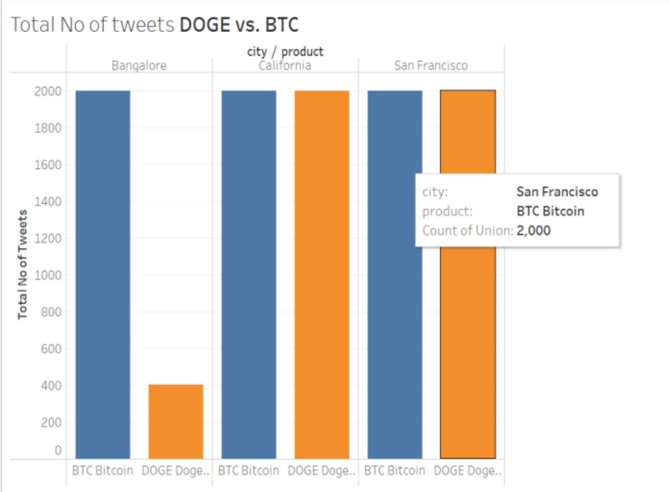
8) Data Gathering and Import: The source files are connected to Tableau for visualisation. The datasets are imported and the union of all the datasets is taken. A Union is a row level connection of tables with same structure.

9) Using tableau to visualize generated dataset: Final step is to create a dashboard using tableau, which is a Data Visualisation tool that is used for obtaining data insights. It helps create graphs and charts which are highly interactive in the form of a dashboard. It provides means for easy analysis of data.

IV. Results

A) Visualisations

1) DOGE vs BTC graph: This graph indicates the number of tweets for both Doge coin and Bitcoin in Bangalore, California City and San Francisco collected over a period of 14 days in December and May. In December 2021 we got



5,500 tweets and in May 2022, we scraped over 10,500 tweets which shows that the popularity of BTC and DOGE is increasing.

Fig. 4 Total No. Of Tweets

2) Original Vs. Duplicate graph:
Some bot accounts tweet the same tweet many times in order to make their products, which in this case is cryptocurrency, more popular and more noticeable on social media platform. So, to analyse the data we need to filter out such tweets which aren't genuine i.e., are duplicates.

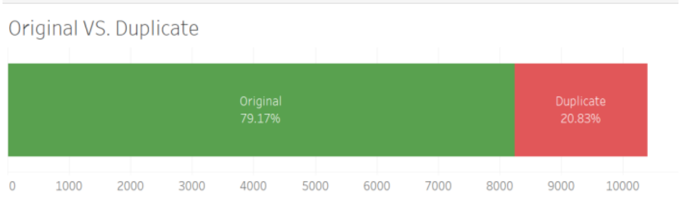


Fig. 5 Ratio between genuine Vs Fake tweets.

Over 21% of the tweets are unreliable i.e., are duplicates. To find out which Cryptocurrency has more reliable i.e., genuine tweets; we have created a city-wise and currency-wise Original Vs Duplicate chart as given below.

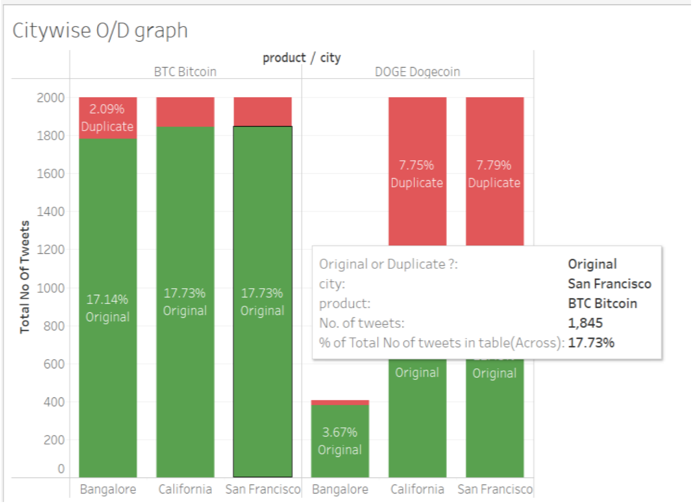


Fig 6. City wise original Vs Duplicate chart

As we can see, in all cities, Dogecoin has more unreliable data i.e., duplicate tweets which are possibly tweeted by crypto spam bots.

3) The Sentiment Score graph: This visualisation shows the number of tweets for a specific sentiment score. It also shows the polarity of tweets as positive, negative or neutral values from the range of -5 to 5

-5 indicates very negative
5 indicates very positive tweets
0 indicates neutral tweets. (Which we have excluded by creating a polarity filter since they don't much contribute to sentiment analysis)

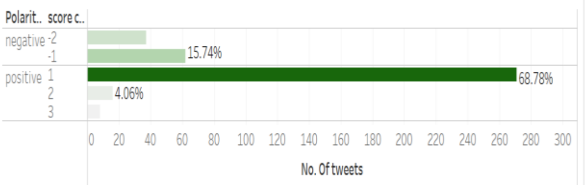


Fig 7. A graph showing sentiment scores when duplicate tweets are included

The original-duplicate filter can be adjusted for the visualisation of original, duplicate or all tweets.

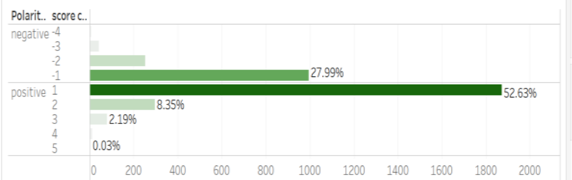


Fig 8. A graph showing sentiment scores when only original tweets are included

Both the charts show that cryptocurrency is having a positive sentiment among more than 50% of the users. Also, on including only original tweets, positve sentiment has decreased to 52.6% from 68%. This proves that the fake twets(duplicates) were indeed spams in order to increase the popularity of cryptocurrency.

4)City-wise Polarity Graph: This graph is plotted from city names and cryptocurrency coins. The Doge coin and the Bitcoin are mentioned with the cities Bangalore, San Francisco and California. This graph shows the polarity of positive and negative tweets for every coin in each of the respective cities.

The positive % indicates how much the coins are liked or popular in that city.

The negative % indicates how much the coins are disliked or are not popular in that city.

Fig 9. Polarity Map

There are more supporters of Bitcoins than Dogecoin i.e., polarity of Bitcoins is positive in comparison to Doge and California has most supporters of bitcoins among the 3 cities which makes sense as California is among the top most cities having the greatest number of cryptocurrency users.

Through these visualisations we get a better insight in the marketing analysis of the cryptocurrency.

5)Time Span Graph:

We have collected tweets from the month of Decemeber 2021 and May 2022 and plotted graph taking date and no. of tweets made as parameters in order to analyse the trade-off between popularity of crypto and its price.

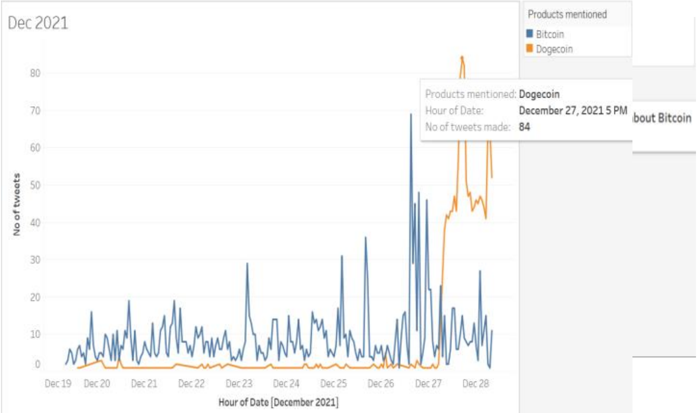


Fig 10. Time Span Graph For December 2021

We can see that Doge is garnering more interest than BTC for few days. Now let's compare it with the price fluctuation chart of Doge in that period: -

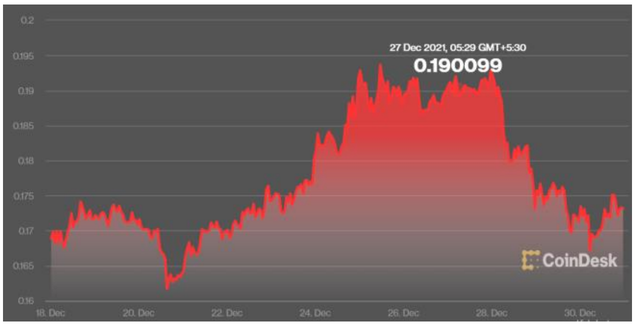


Fig 11. Doge Value Chart For December 2021(Source: coindesk.com)

We can see that when doge had most tweets(even more than BTC), its price increased in the exact time period. This is because everyone was interacting with the tweet by Elon Musk in which he said that Tesla will accept Doge as valid payment method. This resulted in increased interest in his followers to buy Doge which increased its value.

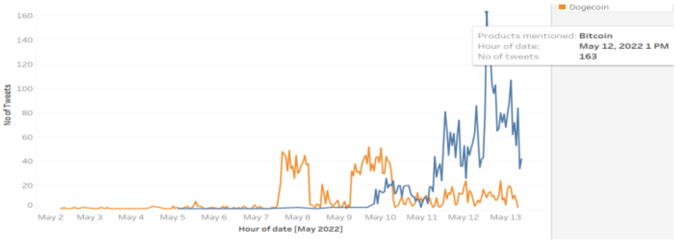


Fig 12. Time Span Graph for May 2022

Fig. 12 shows current i.e. May 2022 time span graph between BTC and Doge.

As we had seen in the past trends, Dogecoin is highly volatile and just in the span of 4 months, it has again become a lot less popular than BTC and its price has also reduced to approximately 8 U.S. cents as of 14 May 2022.

Bitcoin is less volatile than Doge and holds a steady interest of people over the span of 5 months.

V. Conclusion and Future Work

A. Conclusion

We can conclude from this study that the interest of people in cryptocurrency has increased over the past few months. It's rapidly-growing market which can be swayed by sentiments and perceptions of people who express their opinions on top social media platforms like Twitter. If a person is highly influential e.g Elon Musk, only one tweet by him about cryptocurrency is enough to dramatically change its value as we had analyzed in this paper earlier. This effect doesn't remain permanent though as people's interest and sentiments towards a paticular product changes over time but one thing is certain that cryptocurrency market is indeed heavily influenced by social media, especially Dogecoin which was created as a joke intially but mentions from everyone including Musk to Mark Cuban has kept it in the spotlight, and hence its value increased for a short period of time but then again decreased once the interest subsided. Bitcoin is lot less volatile which is holding a steady interest of people online and hence is the most valued cryptocurrency as of May 2022.

It was also derived from the results that number of duplicate tweets affects the polarity of a particular product. This can be either becuase people tend to share positive messages more likely or because bitcoin companies invest a fair amount of time in social media marketing in order to get overall opinion about the currency as positive as possible. Hence, to get a real picture, we need to consider only the number of original tweets.

It was also found from the results that the new sentiment score function improves performance of the standard lexicon-based sentiment analysis algorithm and delivers fast and accurate results.

B. Future Work

Our project is implemented in the English language only. In future if we want to expand and make this a multilingual analysis project then, we can simply translate lexicon files into the desired language using Google Translate. People have used this approach in the past and it provides reasonable results for some languages like Spanish.

Then there are certain Search term and location restrictions. A significant number of tweets are duplicates or retweets. It is the same info over and over again, which is worthless for analysis and should be filtered out quite often. So, we are left with only half the number we have scraped after removing duplicates which decreases our sample size. Quite often with geocoding, we find that there are simply not enough tweets available in the Twitter history. Hence you will get an error message saying you requested X number of tweets, but only a small number could be served. It is actually quite hard to get 2000 tweets for most of the topics when we use geocoding, only super popular keywords can bring that amount on the table. We can use related keywords and fuse the results, or we can use auto-harvesting tools like IFTT (If-This-Then-That).

REFERENCES

[1] Chahat Tandon, Sanjana Revankar, Hemant Palivela, Sidharth Singh Parihar,International Journal of Information Management Data Insights,Volume 1, Issue 2,2021,100035,ISSN 2667-0968,https://doi.org/10.1016/j.jjimei.2021.100035

[2] Shathik, Anvar & Karani, Krishna Prasad. (2020). A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques. 2581-7000. 10.5281/zenodo.3977576.

[3] Liu, Ning & Shen, Bo & Zhang, Zhenjiang & Zhang, Zhiyuan & Mi, Kun. (2019). Attention-based Sentiment Reasoner for aspect-based sentiment analysis. Human- centric Computing and Information Sciences. 9. 10.1186/s13673-019-0196-3.

[4] Gopu, Magesh & Swarnalatha, P.. (2017). Analyzing customer sentiments using machine learning techniques. ijciet. 8. 1829-1842.

[5] Singh, Jaspreet & Singh, Gurvinder & Singh, Rajinder. (2017). Optimization of sentiment analysis using machine learning classifiers. Human-centric Computing and Information Sciences. 7. 10.1186/s13673-017-0116-3.

[6] Mäntylä, Mika & Graziotin, Daniel & Kuutila, Miikka. (2016). The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. Computer Science Review. 27. 10.1016/j.cosrev.2017.10.002.