Predicting Bitcoin prices with Social Media & Google Trends

Aryan Tyagi
(Under the supervision of Dr. Supriya Mohanty. Assistant Professor)
Department of Civil Engineering, Indian Institute of Technology (BHU)
Varanasi, India
aryan.tyagi.civ20@itbhu.ac.in

***Abstract*—Cryptocurrency is a medium of exchange, just like the Indian Rupee, but is digital and uses encryption to control and verify the creation of monetary units and transfer of funds. With the Bitcoin market capitalization over 700 billion dollars, it is increasingly popular on platforms like Twitter which is used as a news source influencing purchase decision of its users. Predicting cryptocurrency price movements is a challenging task due to the highly stochastic nature of the market. Understanding impact of social media on cryptocurrency prices puts any trader in an advantageous position. This paper puts forward an accurate method of predicting Bitcoin price changes using Google Trends, Tweet volume and sentiment analysis. Taking Google Trends data and tweet volume, reasonably accurate predictions of cryptocurrency price changes were possible.**

***Index Terms*—Bitcoin, Cryptocurrency, sentiment analysis, Twitter, Google Trends, correlation coefficient, NLP, market prediction**

I. Introduction

Total cryptocurrency market capitalization, or the value of all cryptocurrencies in existence, peaked in May 2021 at about $2.4 trillion, up from around $200 billion in 2019. 47% of that was taken up by Bitcoin alone. Due ot the significant value of these currencies, many see them as an investment opportunity. More and more investors are joining this market due to high returns. The result is huge swings in the prices of cryptocurrencies. In 1 week, from Feb 2, 2022 to Feb 9, 2022, value of a unit of Bitcoin increased by 20% from INR 27.5L to 33.1L and in the same time from Feb 15 to Feb 21, 2022 it decreased back to INR 27.59L. This kind of volatility means uncertainty for investors thus raising the question of weather we can help these investors in making the right investment decisions. Ladislav Kiroufek found out how Bitcoin is a unique asset, its price behaves in ways similar to both a standard financial asset and a speculative one. [1] Choi and Varian [2] and Ettredge et al. [3] found that web-based search data and Google Trends data could be used to make predictions of macroeconomic statistics such as unemployment rates & automobile sales. Building off of this research we also find a correlation between Google Trends data and fluctuating bitcoin prices, then including tweet volume in a linear model to predict bitcoin prices. Also, we studied the effects of Tweets sentiment and found out how it was not a good measure of predicting bitcoin prices. Results showed that the model was able to predict Bitcoin price movements quite well.

The paper is divided into 7 parts. Section II recalls the background knowledge on Correlation, R^2 score, sentiment analysis & general knowledge about Twitter & Cryptocurrencies. Section III gives detailed view on data collection, preparation. In Section IV we analyze the data, choosing features for model and the correlation of data collected and Bitcoin price movements. Finally Section V details about the Results & Limitations of the model. Section VI discusses about ethical norms to be followed & Section VII summarizes our conclusions and envisions some future work.

II. Background

Knowledge about these topics will be required to properly understand the analysis detailed later in this paper. Also, in this section it will be cleared why Twitter and Google Trend were chosen for this particular task.

A. Bitcoin and other Cryptocurrencies

Cryptocurrency is a digital payment system that doesn't rely on banks to verify transactions. Bitcoin was founded in 2009, developed by "Satoshi Nakamoto" – widely believed to be a pseudonym for an individual or group of people whose precise identity remains unknown.[1] It was the first cryptocurrency and is still most commonly traded. With the launch of Bitcoin, "Satoshi Nakamoto" published a paper "Bitcoin: A Peer-to-Peer Electronic Cash System" [4] which described a peer-to-peer payment system using electronic cash that could be sent directly from one party to another without the use of a third party to validate the transaction. This innovation is created by the use of the "blockchain" which is like a shared ledger on the peer-to-peer network where all transactions are verified by the network so they cannot be forged [4]. Blockchain technology provides security, privacy, and a distributed ledger which makes them applicable for internet-of-things applications, distributed storage systems, healthcare, and more [5]. Cryptocurrencies are tied to the blockchain because they provide the incentive for machines, and the electricity they consume, to run and validate the blockchain. In less than a decade, cryptocurrency has exploded, as of March 2022 there are 18,465 cryptocurrencies in existence indicating its popularity and scale of use.

[1] https://en.wikipedia.org/wiki/Bitcoin

B. Twitter

Founded in 2006, Twitter is an American microblogging, wiz a medium that allows for smaller and more frequent updates than blogging[2] and social networking service on which users post and interact with messages known as "tweets" with a maximum length of 140 characters. In November 2017 the maximum limit was increased to 280 characters.[3] Registered users can post, like and retweet tweets, but unregistered users can only read those that are publicly available. Also, users can add "hashtags" to a tweet, which is the # symbol followed by the text like #bitcoin. Users interact with Twitter through browser or mobile frontend software, or programmatically via its APIs.[4] Highlighting its reach and impact, Twitter has more than 200 million active monetizable users and around 500 million tweets per day, also 83 percent of the world's leaders are on Twitter. One of the example of Twitter's reach and impact was on January 15 of 2009 when a US Airways flight crashed into the Hudson river & the image was posted to Twitter even before the media.[5]

All this makes Twitter an amazing source of data on what people think about any given topic, which is utilised here by collecting text data and exploring the possible relationship between cryptocurrency prices and social media trends.

C. Google Trends

92.01 percent of all internet searches is done on Google search engine[6]. Google Trends is a website by Google that analyses the popularity of top search queries in Google Search across various regions and languages. The website uses graphs to compare the search volume of different queries over time. Google Trends data provides information on how popular given search term is relative to other search term at any given time which can also be compared over time. Like how Simeon Vosen et al. showed how Google Trends outperforms survey-based indicators [6]. Furthermore, it was shown by Tobias Preis et al. that there is a correlation between Google Trends data of company names and transaction volumes of the corresponding stocks on a weekly time scale, [7] which reinforces the argument that Google Trends data impacts the market.

D. Covariance, Correlation & the R^2 score

Variance of a random variable X gives us how much its value deviates from its mean μ_x or $E(X)$ where E(X) is the expectation. Mathematically speaking,

For a Poulation

$$VAR(X) = E(X - \mu_x)^2$$

Similarly, Covariance for two random variables X,Y

$$COV(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

where:

- μ_X or $E(X)$ is the expectation value of X
- μ_Y is the expectation value of Y
- $COV(X, Y)$ is the covariance of X and Y

Pearson's correlation coefficient is commonly represented by the Greek letter $\rho(rho)$.

$$\rho_{X,Y} = \frac{COV(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_X \sigma_Y}$$

$$\mu_X = E[X]$$

$$\mu_Y = E[Y]$$

$$\text{VAR(X)} = \sigma_X^2 = E[(X - E(X)^2] = E[X] - (E[X])^2$$

$$\text{VAR(Y)} = \sigma_Y^2 = E[(Y - E(Y)^2] = E[Y] - (E[Y])^2$$

$$E[(X - \mu_X)(Y - \mu Y)] = E[(X - \mu_X)]E[(Y - \mu Y)] = E[XY] - E[X]E[Y],$$

$$COR(X, Y) = \frac{COV(X, Y)}{\sqrt{VAR(X)VAR(Y)}}$$

So, $\rho_{X,Y} =$

$$\frac{(E[XY] - E[X]E[Y])}{\sqrt{E[X^2] - (E[X]^2)}\sqrt{E[Y^2] - (E[Y]^2)}}$$

For a sample

Pearson's correlation coefficient, when applied to a sample, is commonly represented by $r_{xy}$and may be referred to as the sample correlation coefficient.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

- n is sample size
- x_i, y_i are the individual sample points indexed with i

R-squared score is also known as coefficient of determination of the prediction.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yTrue_i - yPred_i)^2}{\sum_{i=1}^{n}(yTrue_i - \mu_{yTrue})^2}$$

The best possible score is 1.0 and it can be negative. A constant model that always predicts the expected value of y, disregarding the input features, would get a score of 0.0.

E. Sentiment Analysis

Data generated from conversations, declarations or even tweets are examples of unstructured data. Unstructured data doesn't fit neatly into the traditional row and column structure of relational databases, and represent the vast majority of data available in the actual world. It is messy and hard to manipulate. This vast amount of unstructured data has led to the creation of "natural language processing" (NLP) as an area of study and development. NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages.[7]

Further in this paper sentiment analysis is used, sentiment analysis is the act of extracting opinions and emotions that are

[2]https://en.wikipedia.org/wiki/Microblogging

[3]https://developer.twitter.com/en/docs/counting-characters

[4]https://en.wikipedia.org/wiki/Twitter

[5]https://www.brandwatch.com/blog/twitter-stats-and-statistics/

[6]https://www.oberlo.com/statistics/search-engine-market-share

[7]https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

expressed in any text. "VADER" (Valence Aware Dictionary and sEntiment Reasoner) [8] analysis is used in this study since the goal is to classify collected tweets as subjective (positive or negative opinion) or objective (providing just information) tweets. VADER sentiment analysis returns a score between -1 to 1, indicating most negative to most positive sentiment. The score is the sum of sentiment scores of all lexical features and then normalizing it by the following formula-

$$y = x/(\sqrt{x^2 + c})$$

- x = Summation of sentiment scores of all the lexical features
- c = Normalization parameter

VADER sentiment analysis works on lexical features mapped to emotion intensities known as sentiment scores. These sentiment scores are generated by human raters from Amazon Mechanical Turk. VADER analysis also works better with informal texts thus is a better fit for social media text compared to similar sentiment analysis packages such as TextBlob which works better with formal texts.[8]

III. Collecting Data

A. Google Trends Data

Google Trend provides non-real time search data from the year 2004. The problem is Google normalizes search data by returning SVI(search volume index) instead of search volumes. Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics.[9] Zero indicates the lowest relative search interest for the given keyword whereas 100 indicates the maximum search interest within the selected time range.

Data scraping is done using pytrends. An unofficial API for Google Trends, allowing simple interface for automating downloading of reports from Google Trends. Google trends data was collected for the terms "bitcoin", "cryptocurrency" and "crypto", not abbreviations like "btc" because of the ambiguity they pose (existing boot brands named "btc"). Also the data was collected for the time period between April 2016 to Jan 2022. When trends data is queried on Google Trends website for a period of longer than about 3 months, the SVI returned are weekly SVI(aggregated at a weekly level). Pytrend returns weekly SVI values for a time period equal to or less than 5 years, after that it return monthly SVI values.

To combine daily and weekly search data from Google Trends creating a daily time series for over a period longer than the 90 days, method detailed by Erik Johansson is adopted with slight modifications. [9]

- In this, daily SVI values are collected in 3 month increments for the entire time period of interest. Now data is collected for the same time period just aggregated at a weekly level to get the weekly SVI. Then an adjustment factor is calculated. It is the weekly SVI divided by the daily SVI for the first day of the week. Since, Trends only shows data for popular terms, so search terms with low volume appear as "0".[10] For avoiding serious inconsistencies in the data, such SVI values were incremented by 0.5 . This way, the relative volumes between weeks is unchanged and the data is comparable over periods longer than the 90 days provided by Google Trends.

Date	Daily SVI	Weekly SVI	Adjusted values
1.1.2014	72	20	0,277*72 = 20
2.1.2014	96		26,7
3.1.2014	16		4,4
4.1.2014	70		19,4
5.1.2014	61		16,9
6.1.2014	97		26,9
7.1.2014	44		12,2

B. Collecting Tweets with snscrape

To collect Twitter data Tweepy, which is an open-source Python library, can be used for accessing the Twitter API. The reason Tweepy was not used here because of the scraping limit of 3200 tweets which it only allows retrieval of tweets up to 7 days ago, so there is no access to historical data. Luckily, snscrape has stood out as a library that allows one to scrape tweets without the restrictions of Tweepy. Released on July 8, 2020, snscrape is a scraping tool for social networking services (SNS). It scrapes things like users, user profiles, hashtags, searches, threads, list posts and returns the discovered items without using Twitter's API.[11] Using python wrapper for snscrape, around 12,00,000 tweets were collected between a time period of 2 months.

C. Cleaning tweets and sentiment analysis

In this paper, VADER (Valence aware dictionary for sEntiment Reasoner) was used for sentiment analysis. But before the analysis, some pre-processing of tweets is required. The incredible thing about VADER is it doesn't require a great deal of pre-processing to work. Unlike with some supervised methods of NLP, pre-processing necessities such as tokenisation and stemming/lemmatisation are not required. VADER is even smart enough to understand the valence of non-conventional text, including emojis (i.e. :-( ), capitalisation (i.e. sad vs SAD). VADER removes stop words automatically so there is no need to do so ourselves. [12] Remaining Tweet noise such as hashtags and mentions (#, @), URLs and links, all non-alphanumeric characters were removed. Punctuations such as period, comma, exclamation mark, question mark, etc. were also removed as they can cause bias in sentiment analysis.

[8]https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540

[9]https://support.google.com/trends/answer/4365533?hl=en

[10]https://support.google.com/trends/answer/4365533?hl=en

[11]https://github.com/JustAnotherArchivist/snscrape

[12]https://towardsdatascience.com/pre-processing-in-natural-language-machine-learning-898a84b8bd47

D. Collecting Tweet volume

Tweet volumes were scraped from a site[13], which provides tweets per day for "#Bitcoin" dating back to April of 2014.

IV. Data Analysis

A. Tweet Volume

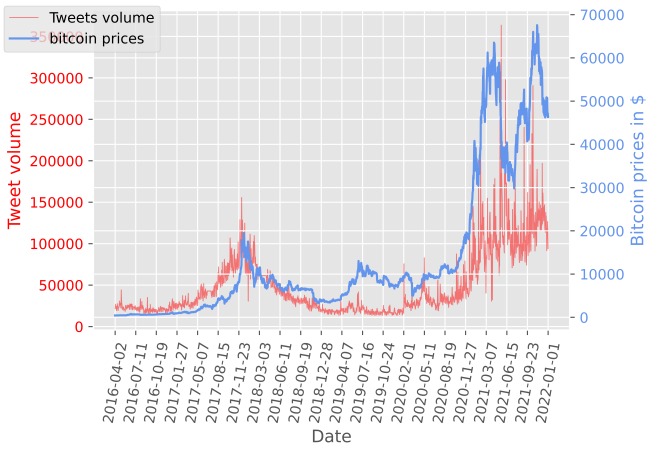

Fig. 1. Daily Bitcoin Tweet Volumes. Plot created in Python using matplotlib.

shows Daily tweet volume and Bitcoin price data from 2016 to 2022. Pearson R coefficient for Tweet volume and Bitcoin prices was 0.7885 while the p-value was 0.0 indicating a strong correlation between two. Thus, it will be considered as an input for the model.

B. Google Trend SVI data for "Bitcoin"

After adjustment in the daily SVI values, it is necessary to check its relationship with bitcoin prices. The Pearson R of the correlation comes out to 0.4064 with p value in

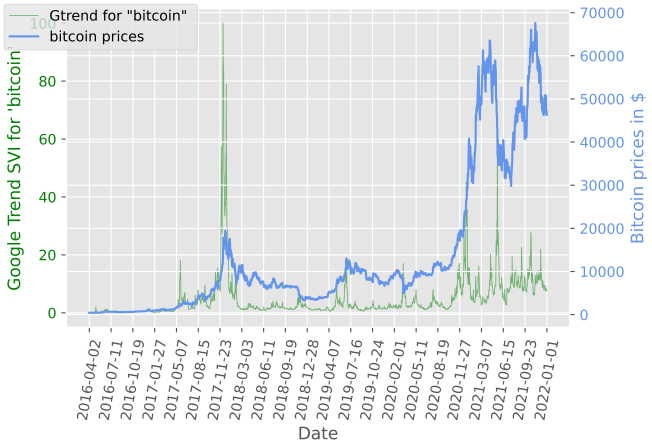

Fig. 2. Adjusted daily Google Trend Search Volume Index for "Bitcoin". Plot created in Python using matplotlib.

the order of e-84 which is negligible. Since Pearson R correlation lies between 0.4 to 0.5, there is moderate correlation between Google Trend SVI data for the term "Bitcoin". Thus, it is established that there is a lack of any strong correlation between these two and it will not be considered as an input to the model.

C. Google Trend SVI data for "Crypto"

Pearson R of the correlation comes out to 0.86098 with p value 0.0 Indicating very strong correlation. As a result, Google Trend data for the term 'crypto' will be used in prediction. From figure 3 too, price is highly correlated with

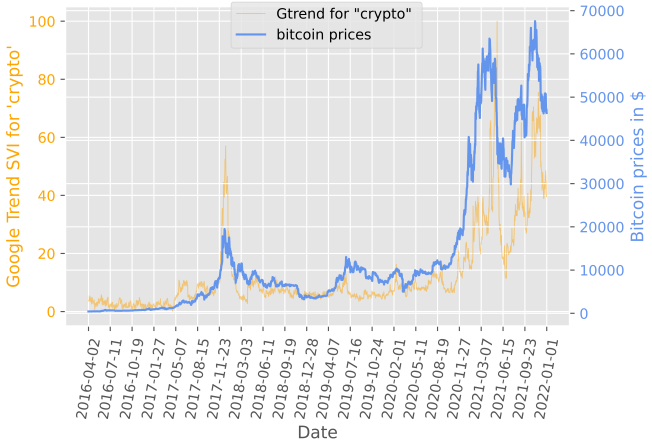

Fig. 3. Adjusted daily Google Trend Search Volume Index for "Crypto". Plot created in Python using matplotlib.

Google Trends data. The pattern holds for increasing as well as decreasing bitcoin prices.

D. Google Trend SVI data for "Cryptocurrency"

Pearson R comes out to 0.5481 and p-value 4.2568e-165. This shows moderate correlation and thus google trend data will be taken as an input to the model. From figure 3 too, it

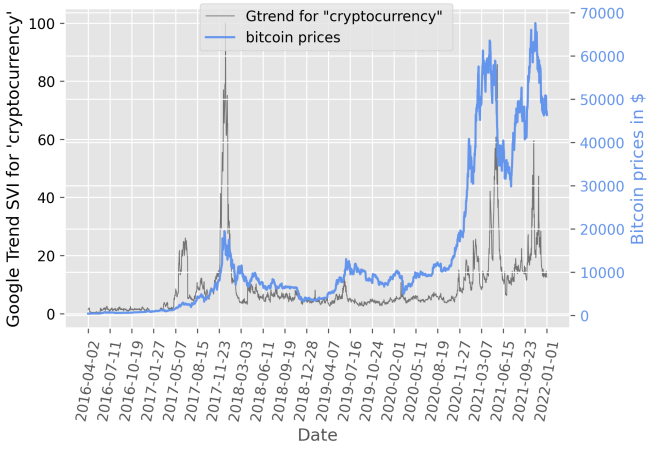

Fig. 4. Adjusted daily Google Trend Search Volume Index for "Crypto". Plot created in Python using matplotlib.

is established that the price is correlated with Google Trends data.

[13]www.bitinfocharts.com

E. Sentiment analysis on Tweets

Twitter estimates that roughly 8.5% of its monthly active users are actually bots, research shows that humans can view Twitter bots as a credible source of information. [11] This shows that Twitter bots posting tweets of either negative or positive sentiment can have significant effects on the decision making, thus impacting cryptocurrency prices.

However certain factors pushed us not to use sentiment analysis for prediction purpose.

1) Overall sentiment of tweets tended to stay positive regardless of how bitcoin prices were changing as shown

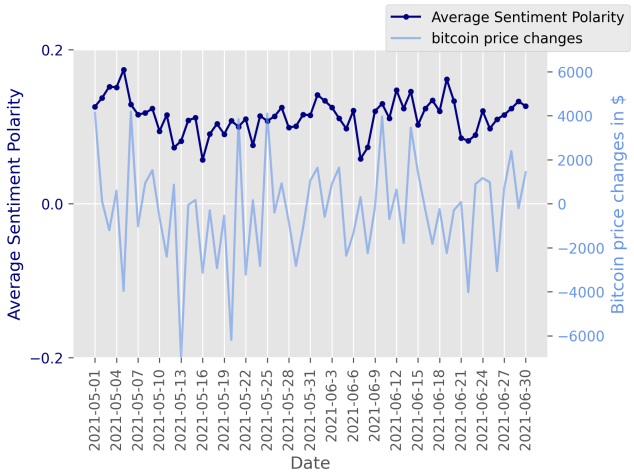

Fig. 5. Fig shows plot containing data between 1st May 2021 and 30th June 2022, on the left y-axis is the average sentiment polarity of daily tweets and on the right y-axis is bitcoin prices changes. Plot created in Python using matplotlib.

in figure 5. This may be due to people tweeting positively as they are interested in bitcoin for its positive attributes (it is decentralized, open source, its low transaction fees etc.)

2) Majority of the tweets did not contain any sentiment as in they were only positing facts or serving the function of advertising. Analysis showed that overall tweets were very broad and generic. Less than half of tweets made on "Bitcoin" were objective, all the other tweets were strictly neutral. This observation also relates to how Hamid Bagheri et al. [12]showed that for most queries, the percentage of the neutral tweets remains significantly high than positive or negative tweets. Even though only half of the tweets collected provide positive/negative sentiment making the tweets overall objective, there is still a chance that the sentiment could be a benefit to the model if any relationship between sentiment and price changes is present. In figure 5 even for falling bitcoin prices, the sentiment of tweets still remains positive and there is a clear lack of relationship between sentiment polarity and prices changes.
3) Close observation to the sentiment data between 1st May 2021 to 30th June 2021, reveals a significant pattern.

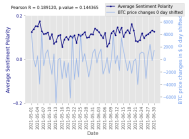

Fig. 6. 0 shift

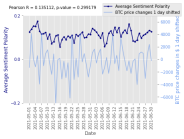

Fig. 7. 1 day shift

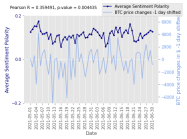

Fig. 8. -1 day shift

The sentiment polarity leads the bitcoin price changes. Implying, if the sentiment polarity of tweets fall/rise, the result is somewhat reflected in the bitcoin prices around half to a day, after the sentiment changes. Figure indicates higher Pearson R coefficient between sentiment polarity and "lagging" bitcoin prices compared to zero shift/leading prices. One potential reason that explains this behaviour is people discussing about cryptocurrency on Twitter before shaping their trading decisions, thus having an effect on other's decisions too. As shown by Java et al., 2007 the most common use of Twitter was to describe what people were doing, which fell into the "daily chatter" category. However, the correlation coefficient is less than 0.5 and therefore it is not as good of a measure.

The analyses mentioned above establishes that Twitter sentiment is not consistent price changes due to the lack of a clear relationship between the sentiment of tweets and prices changes the sentiment analysis will not be used as an input to the model.

All of the data collected for different features was then normalized using *MinMaxScaling* method

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where

- X is the value before min-max scaling
- X_{min} is the minimum value of feature X
- X_{max} is the maximum value of feature X
- X_{scaled} is the min-max scaled value of feature X

V. Results & Limitations

In the end, 3 features were considered as inputs. Google Trends data for "crypto" and "cryptocurrency", Tweets volume were highly correlated with bitcoin prices. Also the relation held during 'ups' as well as 'downs' of bitcoin. Multiple linear regression was selected as the modelling algorithm of choice so as to determine a mathematical relationship among these variables. The overall data set was split into two sections, 89 percent for training the model and 11 percent for testing it and cross validation was not included into the model. The results of the model are outputted in the graph below.

The coefficient of determination or R^2 for training is 0.7985 with test R^2 score being 0.4735. The Pearson R coefficient for prediction and actual prices came out to 0.9395 which indicated very high correlation between prediction and actual values.

Although the model can generally predict cryptocurrency trends it is not quite able to do so on a daily basis. Also, at

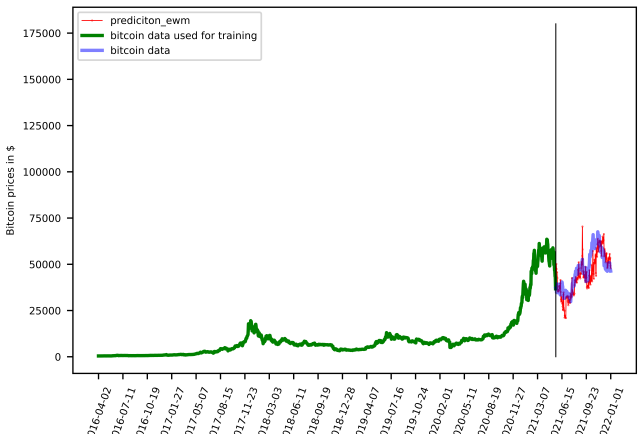

Fig. 9. Actual Bitcoin prices are in blue, training data is in green, test results are in red. Plot created in Python using matplotlib.

some days the model can predict far off values than actual which is due to the nature of data used for training thus unavoidable.

VI. Morality & Norms

Working with Twitter data comes with the moral reponsibility of respecting privacy of users entrusting the data to us. The tweets collected in this study are all made publicly available by the author through sending the tweets through Twitter. Since many people dont realize their tweets are publicly available and accessible through Twitter's API, we chose not to make these tweets publicly available to protect their privacy as much as possible. Other concerns related to cryptocurrencies are that they are currenctly unregulated, this anonymity promotes the exchange of illegal goods and services though cryptocurrency, this makes it immensely difficult to predict the market trends. Also, this model should only be considered for understanding the relationship between social media & cryptocurrency trends. Any false confidence in the effectiveness of this model may lead to weighty loss of wealth. Thus, it is clearly explained how the model works and what are the limitations of it. This study allows people to make better decisions on how they should store their wealth and minimize the risks associated with cryptocurrency. Not understanding the risks associated with cryptocurrency investing, or hiding these risks from others can result in deleterious results like loss of money. This is not an area to be taken lightly, "Let the buyer aware".

VII. Conclusion & Future Works

In this study we have shown how Google Trends and Tweet volume data is able to follow the cryptocurrency price movements & how sentiment of tweets was more correlated with leading bitcoin prices indicating people tweeting before acting on their investment decisions. (However sentiment analysis was not considered as an input given the low correlation between sentiment and price changes). We have shown that the Trend SVI are highly correlated with Bitcoin prices as are Tweet volumes. With these as inputs to a multiple linear regression model, the model fairly accurately predicted future price changes. Additionally time series forecasting/more complex models not just linear models, can also be used further building on our study.

References

[1] Kristoufek, L.: What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. PLOS ONE 10(4) (04 2015) 1-15
[2] HYUNYOUNG, C., HAL, V.: Predicting the present with google trends. Economic Record 88(s1) 2-9
[3] Ettredge, M., Gerdes, J., Karuga, G.: Using web-based search data to predict macroeconomic statistics
[4] Nakamoto, Satoshi. Bitcoin: A Peer-to-Peer Electronic Cash System, 2009.
[5] Miraz, M.H., Ali, M.: Applications of blockchain technology beyond cryptocurrency. CoRR abs/1801.03528 (2018)
[6] Vosen, S. and Schmidt, T. (2011), Forecasting private consumption: survey-based indicators vs. Google trends. J. Forecast., 30: 565-578. https://doi.org/10.1002/for.1213.
[7] Tobias Preis; Daniel Reith; H. Eugene Stanley (2010). "Complex dynamics of our economic life on different scales: insights from search engine query data".
[8] Hutto, C.J. and Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
[9] (2014) Creating daily search volume data from weekly and daily data. https://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html
[10] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[11] Spence, P.R.; Shelton, Ashleigh; Edwards, Chad; Edwards, Autumn (2013). "Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter". Computers in Human Behavior. 33: 372–376. doi:10.1016/j.chb.2013.08.013.
[12] Sentiment analysis of twitter data (2017), Hamid Bagheri, Md Johirul Islam, https://doi.org/10.48550/arXiv.1711.10377