Analyzing The Effect of Social Media Activity on The Performance of Cryptocurrencies

Project Group 6
Archit Patil, Rohan Puthran, Inan Ates, Peyman Alipour

Abstract

Globalization and technology have increasingly created global connectivity. The latter has thus created opportunities and a demand for currencies that can keep up with the global exchange of goods, services, and ideas. Cryptocurrency has been the best choice for filing this need. With the exponential increase in computing power over the past half century and the increased focus on textual methods driven by the requirements of Internet search engines, the application of this technique has permeated most disciplines in one way or another. In this paper we want to investigate the relationship between sentiment of tweets and the cryptocurrency market. To this aim, first we analyzed the sentiment of tweets using Vader, Harvard and LM (Loughran and McDonald) dictionaries and then in the second part we studied the relationship between time series of sentiment analysis and time series of price and volume of Bitcoin and Etherum by using Vector Autoregression(VAR) model. Our results show that there is no significant relationship between various sentiment analysis results and Bitcoin and Ethereum price and volume.

Keywords: Investor sentiment, cryptocurrency, Bitcoin, Ethereum, Vader sentiment, Harvard sentiment, Loughran&McDonald, VAR

Introduction

A cryptocurrency (or "crypto") is a digital currency that can be used to purchase goods and services and is secured by an online ledger and strong cryptography. It's versatility, technology, and growth as assets for various financial purposes makes cryptocurrency a topic worth researching about. Bitcoin was the first cryptocurrency ever created. The creation of Bitcoin is mysterious as it was created by a person or group of people using the name "Satoshi Nakamoto" and released in (Griffin). Along with the launch of Bitcoin "Satoshi Nakamoto" published a paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" which described a peer-to-peer payment system using electronic cash (cryptocurrencies) 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. Blockchain technology provides security, privacy, and a distributed ledger which makes them applicable for internet-of-things applications, distributed storage systems, healthcare, and more.(Miraz, 2018) According to a recent survey, more than 2,300 US businesses accept Bitcoin, one of the most popular cryptocurrencies ("The Business Benefit of Using Cryptocurrency"). Ethereum (ETH) is the second most popular cryptocurrency after Bitcoin(BTC). As the second-largest cryptocurrency by market capitalization (market cap). Both of these tokens are decentralized, meaning that they are not issued or regulated by a central bank or other authority.

In recent times, it is observed that, out of all the factors, social media and the internet are the ones which affect the functioning of these cryptocurrencies the most. Particularly, we are interested in analyzing the effect of social media activities on the performance of cryptocurrencies.

Unstructured text data makes up the majority of data, whether it's in the form of tweets, articles published on the internet, text messages, emails, or other formats. Because of this massive volume of unstructured data, "natural language processing" (NLP) has emerged as a field of research or development. NLP is a set of techniques for analyzing and comprehending text by computers.

In this paper we use a set of natural language processing tools commonly referred to as "sentiment analysis". Sentiment analysis is the act of extracting and measuring the subjective emotions or opinions that are expressed in text.

The goal of this analysis is to apply sentiment analysis on collected tweets so that it can be determined if the tweets are affecting the cryptocurrencies. In this research, we used three different sentiment analysis methods: Vader Sentiment Analysis ,Harvard's General Inquirer and Loughran-McDonald. Further, we compared the results for these methods and took the best result to train our VAR model.

Literature Review

As the data in every sector of finance have grown immensely, text mining has emerged as an important field of research in the domain of finance. More recently, with the exponential increase in computing power over the past half century and the increased focus on textual methods driven by the requirements of Internet search engines, the application of this technique has permeated most disciplines in one way or another. In accounting and finance, the online availability of news articles, earnings conference calls, Securities and Exchange Commission (SEC) filings, and text from social media provide ample fodder for applying the technology (Loughran and Mcdonald 2016).

Li (2010), in a survey of the literature, provides details on earlier manual-based examples of textual analysis, discusses the modern literature by topical area (e.g., information content, earnings quality, market efficiency), and itemizes a prescient list of potential research topics. His conclusions echo a theme of this paper; that is, the literature needs to be less centered on finding ways to apply off-the-shelf textual methods borrowed from highly evolved technologies in computational linguistics and instead be more motivated by hypotheses “closely tied to economic theories” (Li, 2010).

The literature has primarily relied on two distinct word lists to develop dictionaries using the bag-of-words approach: Harvard IV-4 word lists,6 and the Loughran and McDonald (2011) finance specific word list. The Harvard IV-4 word list was created to determine the tone of texts for sociology and psychology literature. This dictionary is further divided into several different categories with the negative and positive word lists getting the largest use in the nance literature (McGurk et al., 2020).

Loughran and McDonald (2011) argue that the Harvard IV-4 word list is not applicable to finance texts and can lead to the misidentification of sentiment. Loughran and McDonald (2011) argue that the Harvard IV-4 word list does not include a number of key finance specific tokens. Further, certain tokens are misclassified as negative when used in a finance context. In fact, Loughran and McDonald (2011) find that around 74 percent of the negative tokens found in the Harvard IV-4 word list are not deemed negative in a finance context. To address these critiques, Loughran and McDonald (2011) create a finance specific word list to accurately identify the tone of 10-K lings. Loughran and McDonald (2011) find sentiment of 10-K lings using their finance specific dictionaries are more correlated with equity returns compared to sentiment of 10-K lings using the Harvard IV-4 word list. Chen et al. (2014) utilize the Loughran and McDonald (2011) dictionary to create an equity specific investor sentiment index from posts and comments from Seeking Alpha (a crowd-sourced - financial market media source) (McDonald & Loughran, 2011).

After that, Chen et al. (2014) find their investor sentiment index is able to predict stock returns up to three months ahead. Jiang et al. (2019) utilize the Loughran and McDonald (2011) dictionary to create a manager sentiment index from annual and quarterly ling, and conference calls. Overall Jiang et al. (2019) find that manager sentiment is able to predict stock returns using out-sample forecast evaluation.

Considering previous research, in this paper we want to investigate the relationship between sentiment of tweets and the cryptocurrency market. To this aim, first we analyzed the sentiment of tweets using Vader, Harvard and LM (Loughran and McDonald) dictionaries and then in the second part we studied the relationship between time series of sentiment analysis and time series of price and volume of Bitcoin and Etherum.

Objectives and Expected Contributions

By scraping a dataset through Twitter and using historic cryptocurrency prices, we plan to answer these research questions:

1.Does the nature of statements (positive, negative or neutral) obtained from social media affect the price trend of cryptocurrency?

One of our goals in this project is to see whether the tweets related to cryptocurrencies affect the price trends of crypto or not?

We want to analyze if a positive or negative tweet leads to an increase or decrease respectively in the price of the crypto or not?

2.Can we predict future cryptocurrency price trends based on current social media presence?

Our goal for this question would be to try and predict whether the crypto has an upward or downward trend on the basis of all the previous trends observed for the particular cryptocurrency.

We expect this work can make the following contributions:

After the successful implementation of the project, we should be able to analyze the effect of social media on the performance of a crypto.

Also, we could predict the price trends of a cryptocurrency based on previous sentiment analysis data.

Methodology

Two cryptocurrencies are taken into consideration, ‘Bitcoin’ and ‘Ethereum’. Snscrape is used to scrape tweets related to these cryptocurrencies on twitter using the keyword ‘Bitcoin’ and ‘Ethereum’. Historical price data in dollars along with volume which means the daily volume of trades in dollars is taken for the two crypto currencies from coindesk.com. Both the tweets and price data is taken for the same time period on which analysis is to be performed. Next, preprocessing is done on tweets to remove stop words and punctuation and to tokenize and lemmatize them. After preprocessing, sentiment analysis is performed on the cleaned and tokenized tweets. Three sentiment analyzers which are Harvard General Inquirer, Loughran-McDonald Master Dictionary and VADER are implemented. Tweets are aggregated by granularity of per day after getting the sentiment scores. We have multiple time-series which include sentiment scores, historic price and volume, all having granularity of a day. VAR(Vector Autoregressive) is implemented since we have multiple time-series varying over time and we want to record the relationship between them.

Data Scraping

Data scraping is done from twitter for our dataset, i.e. cryptocurrency related tweets on which sentiment analysis can be performed. Snscrape is used to scrape tweets related to the two currencies using ‘Bitcoin’ and ‘Ethereum’ as keywords. Data is collected for the dates from 2021-09-01 until 2021-11-30. For each day 500 tweets are taken which makes the entire dataset for each cryptocurrency around 45000 rows. Equivalent historic prices and volume data for respective cryptocurrencies are also taken from Coindesk.com. We plan to aggregate time phases by each day i.e., sentiment analysis will be aggregated for an entire day and then vector autoregression is implemented on it.

Data-preprocessing

Next, data-preprocessing is done which includes the following steps-

- Removed Signs- Removing of @, blank space and hashtags (#)
- Removed Punctuations
- Eliminating stop words
- Tokenization- Using nltk.word_tokenizer()

- Lemmatization- Using WordNet Lemmatizer

We have analyzed tweets and divided them into positive and negative sentiments using sentiment analysis and then by implementation of vector autoregression (VAR) model, we have surveyed whether there is any significant relation between tweets, price and volume trends and try to answer our research questions

Exploratory Data Analysis

- We have two datasets for cryptocurrency i.e., for Bitcoin and Ethereum. Tweets are collected from 2021-09-01 until 2021-11-30. For each day 500 tweets are collected.
- Both have around 45000 rows with 10 attributes which are- day, url, date, content, id , Username, outlinks, outlinksss, tcoutlinks and tcoutlinksss.
- After preprocessing is done, day, url, id, Username, outlinks, outlinksss, tcoutlinks and tcoutlinksss are removed since they are not significant attributes for our analysis and a new attribute(column) called 'Cleantext_lemmatized' is added which has essentially originated from 'content' attribute after performing preprocessing, tokenization and lemmatization.

Final Data Information for Bitcoin

	Unnamed: 0	date	Cleantext_lemmatized
0	0	2021-09-01 23:59:58+00:00	genesis digital asset ซอเครองขด bitcoin เพมอก ...
1	1	2021-09-01 23:59:57+00:00	computer basically useless bitcoin
2	2	2021-09-01 23:59:55+00:00	investing 635 optimal crypto portfolio 365 cas...
3	3	2021-09-01 23:59:54+00:00	right track bitcoin 250000 billionaire tim draper
4	4	2021-09-01 23:59:54+00:00	bitcoin price index usd eur cny gbp rub
...	...	...	...
43995	43995	2021-11-29 23:49:00+00:00	also ist bitcoin ein reines spekulationsobjekt...
43996	43996	2021-11-29 23:49:00+00:00	poke patrick beat saying bitcoin sound money h...
43997	43997	2021-11-29 23:49:00+00:00	turkish subtitle added bitcoin generational we...
43998	43998	2021-11-29 23:48:57+00:00	na binance gente consegue transferir bitcoin p...
43999	43999	2021-11-29 23:48:57+00:00	im saying federate rather centrally controlled...

Final Data Information for Ethereum

	Unnamed: 0	date	Cleantext_lemmatized
0	0	2021-09-01 23:59:57+00:00	ethereum crypto digital currency artist sellin...
1	1	2021-09-01 23:59:56+00:00	gas used currently pending transaction 15 seco...
2	2	2021-09-01 23:59:56+00:00	ethereum coin best riskadjusted return past 24...
3	3	2021-09-01 23:59:54+00:00	best cryptocurrency riskadjusted return past 2...
4	4	2021-09-01 23:59:50+00:00	bgan 4821 bought 10ξ 376765 usd big mullet ene...
...	...	...	...
43995	43995	2021-11-29 23:22:09+00:00	warum ich altcoins investiere weil ich meine g...
43996	43996	2021-11-29 23:22:08+00:00	emily kishimotoinu kishimoto inu kishininjas e...
43997	43997	2021-11-29 23:22:03+00:00	think classic classicswap cl cl etc ethereumcl...
43998	43998	2021-11-29 23:22:03+00:00	ethereum eth current price 444120 1h 036 24h 3...
43999	43999	2021-11-29 23:21:59+00:00	global sdx live grand ambition michele liminab...

Harvard General Inquirer

General Inquirer development has been supported by grants from the USA National Science Foundation and Research Grant Councils of Great Britainand Australia. Until the mid-1990's, it only operated on large mainframe IBM computers that supported the PL/1 programming language. However with its reprogramming supported by the Gallup Organization, first in TrueBasic by Philip Stone and then in Java by Vanja Buvac, the system now provides English-language content analysis capabilities using both the "Harvard" and "Lasswell" general-purpose dictionaries as well as any dictionary categories developed by the user. With today's PC's or Macs, the system, including its disambiguation routines for high-frequency English homographs, usually processes text files on the order of a million words an hour. However, it is not packaged to be commercially available, nor do we intend to commit ourselves to providing the support services such availability would require (*Descriptions of Inquirer Categories and Use of Inquirer Dictionaries*).

Loughran-McDonald Master Dictionary

As an artifact of hackers needing word lists to crack passwords, a variety of word lists are available on the internet. Word lists including proper nouns and abbreviations can exceed 600,000 tokens (a token is a collection of characters).

The Master Dictionary that is used is based on release 4.0 of the 2of12inf dictionary documented at:http://wordlist.sourceforge.net/12dicts-readme.html. The 2of12inf dictionary includes word inflections but does not include abbreviations, acronyms, or names. We use inflections instead of stemming because, in our opinion, especially if the focus is on tone, using explicit inflections is less error prone than extending a word using stemming (root morpheme + derivational morphemes). The 2of12inf word list contains more than 80,000 words. The one-letter words “A” and “I” are not included in our Master Dictionary for two reasons. First, they are not critical content words, and second they are more likely to indicate headers in financial documents.2 All tokens we identify as words contain two or more characters.

The sentiment categories are negative, positive, uncertainty, litigious, strong modal, weak modal, constraining, and complexity. Membership in a given classification is flagged in the corresponding column by the year in which the word was added to that sentiment group. A "negative" year indicates the year when a word was removed from the group (*Resources // Software Repository for Accounting and Finance // University of Notre Dame*).

VADER (**Valence Aware Dictionary for Sentiment Reasoning**)

VADER is primarily used for doing sentiment analysis on social media data. Since our project does analysis of twitter data, VADER is a good option for performing sentiment analysis. It is a lexicon and rule based sentiment analyzer. It gives the polarity of a given text i.e. whether the text is positive or negative. Also, it gives the strength of how much positive or negative sentiment the text has. It uses a dictionary full of lexicons which are used to give the given text a sentiment score. Hence, there is no need to have labels for the dataset to train it since there is already a dictionary of labels which VADER uses.

Vector Autoregression

The vector autoregressive (VAR) model is a workhouse multivariate time series model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.

The vector autoregressive model of order 1, denoted as VAR(1), is as follows:

$$x_{t,1} = \alpha_1 + \phi_{11} x_{t-1,1} + \phi_{12} x_{t-1,2} + \phi_{13} x_{t-1,3} + w_{t,1}$$

$$x_{t,2} = \alpha_2 + \phi_{21} x_{t-1,1} + \phi_{22} x_{t-1,2} + \phi_{23} x_{t-1,3} + w_{t,2}$$

$$x_{t,3} = \alpha_3 + \phi_{31} x_{t-1,1} + \phi_{32} x_{t-1,2} + \phi_{33} x_{t-1,3} + w_{t,3}$$

VAR models are traditionally widely used in finance and econometrics because they offer a framework for accomplishing important modeling goals, including Data description, Forecasting, Structural inference and Policy analysis.

In order to reach a reliable output of VAR, a common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

To apply a VAR model we need to know how many lag we should consider in our model. The selection of lag lengths in AR can sometimes be guided by economic theory. However, there are statistical methods that are helpful to determine how many lags should be included as regressors. In general, too many lags inflate the standard errors of coefficient estimates and thus imply an increase in the forecast error while omitting lags that should be included in the model may result in an estimation bias.

In this paper, to find the order of our VAR model we used Akaike information criterion (AIC). To circumvent the issue of producing too large models, one may choose the lag order that minimizes Akaike information criteria (Tsay, 2005):

$$AIC(p) = \log\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{2}{T}$$

RESULTS

Sentiment Analysis

For our project, we have used three sentiment analysis techniques, so the results for these techniques are:

VADER (Valence Aware Dictionary and Sentiment Reasoner)

These are sentiment results using VADER sentiment analysis for both cryptocurrencies:

	sentiment	Cleantext_lemmatized
1	Neutral	25030
2	Positive	13435
0	Negative	5535

Bitcoin Results

	sentiment	Cleantext_lemmatized
1	Neutral	26872
2	Positive	12475
0	Negative	4653

Ethereum Results

Harvard's General Inquirer

These are the results using Harvard's General Inquirer for both cryptocurrencies:

	Positive_Har_Bit	Negative_Har_Bit	Polarity_Har_Bit	Subjectivity_Har_Bit
0	1.0	0.0	0.999999	0.333333
1	1.0	1.0	0.000000	0.500000
2	2.0	0.0	1.000000	0.200000
3	1.0	0.0	0.999999	0.250000
4	0.0	0.0	0.000000	0.000000
...	...	...	...	...
43995	1.0	0.0	0.999999	0.031250
43996	1.0	3.0	-0.500000	0.333333
43997	1.0	0.0	0.999999	0.111111
43998	0.0	0.0	0.000000	0.000000
43999	5.0	0.0	1.000000	0.227273

Bitcoin Results

	Positive_Har_Eth	Negative_Har_Eth	Polarity_Har_Eth	Subjectivity_Har_Eth
0	1.0	0.0	0.999999	0.100000
1	1.0	0.0	0.999999	0.050000
2	1.0	0.0	0.999999	0.166667
3	1.0	0.0	0.999999	0.100000
4	0.0	0.0	0.000000	0.000000
...	...	...	...	...
43995	0.0	0.0	0.000000	0.000000
43996	0.0	0.0	0.000000	0.000000
43997	1.0	0.0	0.999999	0.062500
43998	0.0	0.0	0.000000	0.000000
43999	0.0	0.0	0.000000	0.000000

Ethereum Results

Loughran-McDonald

These are the results using Loughran-McDonald sentiment method for both cryptocurrencies:

	Positive_lm_Bit	Negative_lm_Bit	Polarity_lm_Bit	Subjectivity_lm_Bit
0	0.0	0.0	0.000000	0.000000
1	0.0	0.0	0.000000	0.000000
2	0.0	0.0	0.000000	0.000000
3	0.0	0.0	0.000000	0.000000
4	0.0	0.0	0.000000	0.000000
...	...	...	...	...
43995	1.0	0.0	0.999999	0.031250
43996	0.0	2.0	-1.000000	0.166667
43997	0.0	0.0	0.000000	0.000000
43998	0.0	0.0	0.000000	0.000000
43999	1.0	0.0	0.999999	0.045455

Bitcoin Results

	Positive_lm_Eth	Negative_lm_Eth	Polarity_lm_Eth	Subjectivity_lm_Eth
0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0
...	...	...	...	...
43995	0.0	0.0	0.0	0.0
43996	0.0	0.0	0.0	0.0
43997	0.0	0.0	0.0	0.0
43998	0.0	0.0	0.0	0.0
43999	0.0	0.0	0.0	0.0

Ethereum Results

Time Series

This part of our project deals with time series formation using cryptocurrency prices and the results obtained from sentiment analysis.

The final time series includes date, sentiment score, closing price, returns and volume for both cryptocurrencies. This is the final time-series we obtained after aggregating data:

	Date	score	Close	Bit_retrrn	Volume
0	9/1/2021	0.3612	48847.02734	0.035006	3.913940e+10
1	9/2/2021	0.0000	49327.72266	0.009793	3.950807e+10
2	9/3/2021	0.0000	50025.37500	0.014044	4.320618e+10
3	9/4/2021	-0.6249	49944.62500	-0.001615	3.747133e+10
4	9/5/2021	0.0000	51753.41016	0.035575	3.032268e+10
...	...	...	...	...	...
83	11/25/2021	-0.1280	57274.67969	0.017512	3.428402e+10
84	11/26/2021	0.8316	53569.76563	-0.066874	4.181075e+10
85	11/27/2021	0.1027	54815.07813	0.022980	3.056086e+10
86	11/28/2021	0.0000	57248.45703	0.043435	2.811689e+10
87	11/29/2021	0.1027	57806.56641	0.009702	3.237084e+10

Bitcoin Time Series

Date	score	Close	Eth_return	Volume
9/1/2021	0.4019	3834.828125	0.110477	3.007089e+10
9/2/2021	0.6369	3790.989990	-0.011497	2.438740e+10
9/3/2021	0.0000	3940.614746	0.038710	2.620777e+10
9/4/2021	0.0000	3887.828369	-0.013486	2.080696e+10
9/5/2021	-0.5719	3952.133545	0.016405	1.837147e+10
...	...	...	...	...
11/25/2021	0.0000	4274.743164	0.008165	1.870536e+10
11/26/2021	0.0000	4030.908936	-0.058732	2.628180e+10
11/27/2021	0.0000	4096.912109	0.016242	1.651569e+10
11/28/2021	-0.4019	4294.453613	0.047091	1.595313e+10
11/29/2021	0.4588	4445.104980	0.034479	1.908648e+10

Ethereum Time Series

VAR (Vector AutoRegression)

This part of our project involves analyzing the time series and obtaining the results for the VAR model. There are various steps involved before running the VAR model, so the results for all the steps are also mentioned below:

Checking stationarity for both time series using results of all sentiment analysis methods

Stationarity check for Bitcoin time series using Harvard's General Inquirer results:

Bitcoin return time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -7.966448e+00
p-value                 2.856477e-12
# lags used             1.000000e+00
# observations          8.600000e+01
critical value (1%)    -3.508783e+00
critical value (5%)    -2.895784e+00
critical value (10%)   -2.585038e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin volume time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -5.357471
p-value                  0.000004
# lags used              3.000000
# observations          84.000000
critical value (1%)     -3.510712
critical value (5%)     -2.896616
critical value (10%)    -2.585482
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Positive time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -9.071912e+00
p-value                 4.279510e-15
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Negative time series

```
ADF test statistic     -9.354055e+00
p-value                 8.149060e-16
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Polarity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -6.855801e+00
p-value                 1.649723e-09
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Subjectivity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -8.681716e+00
p-value                 4.268366e-14
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Stationarity check for Bitcoin time series using Loughran-McDonald results:

Bitcoin return time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -7.966448e+00
p-value                 2.856477e-12
# lags used             1.000000e+00
# observations          8.600000e+01
critical value (1%)    -3.508783e+00
critical value (5%)    -2.895784e+00
critical value (10%)   -2.585038e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin volume time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -5.357471
p-value                  0.000004
# lags used              3.000000
# observations          84.000000
critical value (1%)     -3.510712
critical value (5%)     -2.896616
critical value (10%)    -2.585482
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin LM Positive time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -9.071912e+00
p-value                 4.279510e-15
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin LM Negative time series

```
ADF test statistic     -9.354055e+00
p-value                 8.149060e-16
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin LM Polarity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -6.855801e+00
p-value                 1.649723e-09
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin LM Subjectivity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -8.681716e+00
p-value                 4.268366e-14
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Stationarity check for Ethereum time series using Harvard's General Inquirer results:

Ethereum return time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -7.966448e+00
p-value                 2.856477e-12
# lags used             1.000000e+00
# observations          8.600000e+01
critical value (1%)    -3.508783e+00
critical value (5%)    -2.895784e+00
critical value (10%)   -2.585038e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Ethereum volume time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -5.357471
p-value                  0.000004
# lags used              3.000000
# observations          84.000000
critical value (1%)     -3.510712
critical value (5%)     -2.896616
critical value (10%)    -2.585482
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Ethereum Harvard Positive time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -9.071912e+00
p-value                 4.279510e-15
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Ethereum Harvard Negative time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -2.590939
p-value                  0.094884
# lags used              4.000000
# observations          83.000000
critical value (1%)     -3.511712
critical value (5%)     -2.897048
critical value (10%)    -2.585713
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Ethereum Harvard Polarity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -1.927852
p-value                  0.319074
# lags used              7.000000
# observations          80.000000
critical value (1%)     -3.514869
critical value (5%)     -2.898409
critical value (10%)    -2.586439
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Ethereum Harvard Subjectivity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -2.133462
p-value                  0.231274
# lags used              4.000000
# observations          83.000000
critical value (1%)     -3.511712
critical value (5%)     -2.897048
critical value (10%)    -2.585713
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Stationarity check for Ethereum time series using Loughran-McDonald results:

Bitcoin return time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -7.966448e+00
p-value                 2.856477e-12
# lags used             1.000000e+00
# observations          8.600000e+01
critical value (1%)    -3.508783e+00
critical value (5%)    -2.895784e+00
critical value (10%)   -2.585038e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin volume time series

```
Augmented Dickey-Fuller Test:
ADF test statistic      -5.357471
p-value                  0.000004
# lags used              3.000000
# observations          84.000000
critical value (1%)     -3.510712
critical value (5%)     -2.896616
critical value (10%)    -2.585482
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Positive time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -9.071912e+00
p-value                 4.279510e-15
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Negative time series

```
ADF test statistic     -9.354055e+00
p-value                 8.149060e-16
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Polarity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -6.855801e+00
p-value                 1.649723e-09
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Bitcoin Harvard Subjectivity time series

```
Augmented Dickey-Fuller Test:
ADF test statistic     -8.681716e+00
p-value                 4.268366e-14
# lags used             0.000000e+00
# observations          8.700000e+01
critical value (1%)    -3.507853e+00
critical value (5%)    -2.895382e+00
critical value (10%)   -2.584824e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Choosing The Order for VAR

In this part, we have mentioned the process of choosing the order for VAR model by comparing AIC values of each order:

Bitcoin Harvard	Bitcoin LM	Ethereum Harvard	Ethereum LM
Order = 1 AIC: 11.213565069590262	Order = 1 AIC: 12.623249894047625	Order = 1 AIC: 11.213565069590262	Order = 1 AIC: 12.623249894047625
Order = 2 AIC: 11.483064822910013	Order = 2 AIC: 12.623249894047625	Order = 2 AIC: 11.483064822910013	Order = 2 AIC: 12.623249894047625
Order = 3 AIC: 12.085883071223611	Order = 3 AIC: 12.623249894047625	Order = 3 AIC: 12.085883071223611	Order = 3 AIC: 12.623249894047625
Order = 4 AIC: 12.470018792093143	Order = 4 AIC: 12.623249894047625	Order = 4 AIC: 12.470018792093143	Order = 4 AIC: 12.623249894047625
Order = 5 AIC: 12.623249894047625	Order = 5 AIC: 12.623249894047625	Order = 5 AIC: 12.623249894047625	Order = 5 AIC: 12.623249894047625

Final VAR Model Results

In the VAR model, we obtain equations for each variable in the time series. So, these were the final results for VAR model obtained by us for this project:

Bitcoin Results

Results for equation Bit_return

	coefficient	std. error	t-stat	prob
const	-0.043227	0.063916	-0.676	0.499
L1.Bit_return	0.041814	0.111980	0.373	0.709
L1.Volume	0.000000	0.000000	0.039	0.969
L1.Positive_Har_Bit	-0.007480	0.089287	-0.084	0.933
L1.Negative_Har_Bit	0.166998	0.131043	1.274	0.203
L1.Polarity_Har_Bit	0.002034	0.202470	0.010	0.992
L1.Subjectivity_Har_Bit	-0.224355	0.548918	-0.409	0.683

Results for equation Volume

	coefficient	std. error	t-stat	prob
const	25658946329.289799	11084613575.477999	2.315	0.021
L1.Bit_return	-12273988075.592810	19420098317.677277	-0.632	0.527
L1.Volume	0.312121	0.107900	2.893	0.004
L1.Positive_Har_Bit	-6784151231.178294	15484530712.315434	-0.438	0.661
L1.Negative_Har_Bit	-27094355080.761982	22726013684.923401	-1.192	0.233
L1.Polarity_Har_Bit	3695099344.851302	35113240914.591949	0.105	0.916
L1.Subjectivity_Har_Bit	132313720287.804535	95195830056.517563	1.390	0.165

Results for equation Positive_Har_Bit

	coefficient	std. error	t-stat	prob
const	0.824123	0.177942	4.631	0.000
L1.Bit_return	0.070643	0.311752	0.227	0.821
L1.Volume	0.000000	0.000000	2.037	0.042
L1.Positive_Har_Bit	0.043443	0.248574	0.175	0.861
L1.Negative_Har_Bit	-0.298666	0.364822	-0.819	0.413
L1.Polarity_Har_Bit	-0.067724	0.563674	-0.120	0.904
L1.Subjectivity_Har_Bit	1.862921	1.528183	1.219	0.223

Results for equation Negative_Har_Bit

	coefficient	std. error	t-stat	prob
const	0.559559	0.106139	5.272	0.000
L1.Bit_return	-0.121823	0.185954	-0.655	0.512
L1.Volume	0.000000	0.000000	1.178	0.239
L1.Positive_Har_Bit	-0.051436	0.148270	-0.347	0.729
L1.Negative_Har_Bit	0.083989	0.217609	0.386	0.700
L1.Polarity_Har_Bit	-0.062034	0.336221	-0.185	0.854
L1.Subjectivity_Har_Bit	-0.376766	0.911532	-0.413	0.679

Results for equation Polarity_Har_Bit

	coefficient	std. error	t-stat	prob
const	0.073355	0.073054	1.004	0.315
L1.Bit_return	0.018648	0.127989	0.146	0.884
L1.Volume	0.000000	0.000000	1.218	0.223
L1.Positive_Har_Bit	0.003470	0.102051	0.034	0.973
L1.Negative_Har_Bit	-0.149331	0.149777	-0.997	0.319
L1.Polarity_Har_Bit	0.198534	0.231415	0.858	0.391
L1.Subjectivity_Har_Bit	1.002855	0.627392	1.598	0.110

Results for equation Subjectivity_Har_Bit

	coefficient	std. error	t-stat	prob
const	0.132599	0.018575	7.139	0.000
L1.Bit_return	-0.026123	0.032543	-0.803	0.422
L1.Volume	0.000000	0.000000	1.721	0.085
L1.Positive_Har_Bit	0.001375	0.025948	0.053	0.958
L1.Negative_Har_Bit	-0.001267	0.038083	-0.033	0.973
L1.Polarity_Har_Bit	-0.008382	0.058841	-0.142	0.887
L1.Subjectivity_Har_Bit	0.059720	0.159525	0.374	0.708

Ethereum Results

Results for equation Eth_return

	coefficient	std. error	t-stat	prob
const	-0.008239	0.060822	-0.135	0.892
L1.Eth_return	-0.003011	0.111437	-0.027	0.978
L1.Volume	0.000000	0.000000	0.968	0.333
L1.Positive_Har_Eth	-0.006218	0.115608	-0.054	0.957
L1.Negative_Har_Eth	0.098315	0.154575	0.636	0.525
L1.Polarity_Har_Eth	-0.028208	0.219211	-0.129	0.898
L1.Subjectivity_Har_Eth	-0.390807	0.742856	-0.526	0.599

Results for equation Volume

	coefficient	std. error	t-stat	prob
const	11394114401.450384	5609189623.554939	2.031	0.042
L1.Eth_return	-5507457108.226655	10277044908.254744	-0.536	0.592
L1.Volume	0.391412	0.106145	3.688	0.000
L1.Positive_Har_Eth	5981064239.300002	10661705141.460880	0.561	0.575
L1.Negative_Har_Eth	-7402437011.808105	14255412457.761545	-0.519	0.604
L1.Polarity_Har_Eth	-20899689805.725441	20216389999.574188	-1.034	0.301
L1.Subjectivity_Har_Eth	18281784517.799839	68508558390.861389	0.267	0.790

Results for equation Positive_Har_Eth

	coefficient	std. error	t-stat	prob
const	0.822305	0.133225	6.172	0.000
L1.Eth_return	-0.128668	0.244091	-0.527	0.598
L1.Volume	0.000000	0.000000	0.855	0.392
L1.Positive_Har_Eth	0.327810	0.253228	1.295	0.195
L1.Negative_Har_Eth	-0.357503	0.338582	-1.056	0.291
L1.Polarity_Har_Eth	-0.601828	0.480162	-1.253	0.210
L1.Subjectivity_Har_Eth	0.772632	1.627156	0.475	0.635

Results for equation Negative_Har_Eth

	coefficient	std. error	t-stat	prob
const	0.419701	0.077527	5.414	0.000
L1.Eth_return	-0.164569	0.142044	-1.159	0.247
L1.Volume	0.000000	0.000000	0.518	0.604
L1.Positive_Har_Eth	-0.004084	0.147360	-0.028	0.978
L1.Negative_Har_Eth	0.156152	0.197031	0.793	0.428
L1.Polarity_Har_Eth	0.149943	0.279420	0.537	0.592
L1.Subjectivity_Har_Eth	-0.354265	0.946889	-0.374	0.708

Results for equation Polarity_Har_Eth

	coefficient	std. error	t-stat	prob
const	0.173040	0.064709	2.674	0.007
L1.Eth_return	-0.087170	0.118558	-0.735	0.462
L1.Volume	0.000000	0.000000	0.797	0.425
L1.Positive_Har_Eth	0.100063	0.122995	0.814	0.416
L1.Negative_Har_Eth	-0.283962	0.164453	-1.727	0.084
L1.Polarity_Har_Eth	-0.282768	0.233220	-1.212	0.225
L1.Subjectivity_Har_Eth	0.779168	0.790326	0.986	0.324

Results for equation Subjectivity_Har_Eth

	coefficient	std. error	t-stat	prob
const	0.105505	0.014220	7.419	0.000
L1.Eth_return	-0.024761	0.026054	-0.950	0.342
L1.Volume	0.000000	0.000000	2.512	0.012
L1.Positive_Har_Eth	0.006359	0.027029	0.235	0.814
L1.Negative_Har_Eth	-0.065392	0.036139	-1.809	0.070
L1.Polarity_Har_Eth	-0.073611	0.051251	-1.436	0.151
L1.Subjectivity_Har_Eth	0.368487	0.173679	2.122	0.034

Conclusion

In this paper we want to investigate the relationship between sentiment of tweets and the cryptocurrency market. So, first we analyzed the sentiment of tweets using Vader, Harvard and LM (Loughran and McDonald) dictionaries and then in the second part we studied the relationship between time series of sentiment analysis and time series of price and volume of Bitcoin and Etherum by using VAR.

Our results show that just a few coefficients of the equations are significant and most of them do not have a remarkable impact on the dependent variable.

So in conclusion we can say that, our results showed that there is no significant relation between Bitcoin price and sentiment analysis result. Also there is no significant relation between volume of trade of Bitcoin and sentiment analysis results.

In a similar way, the results for ethereum, we could not find any meaningful relation between ethereum price and sentiment analysis results. Also it is shown that volume of trade of ethereum is not affected by sentiment analysis results.

As the results do not show any significant relationship between sentiment analysis and our desired cryptocurrencies so our model can not be used for predicting future price trends of Bitcoin and Ethereum.

References

The Business Benefit of Using Cryptocurrency. (n.d.). Deloitte. Retrieved December 12, 2021, from https://www2.deloitte.com/us/en/pages/audit/articles/corporates-using-crypto.html

The Business Benefit of Using Cryptocurrency. (n.d.). Deloitte. Retrieved December 12, 2021, from https://www2.deloitte.com/us/en/pages/audit/articles/corporates-using-crypto.html

Descriptions of Inquirer Categories and Use of Inquirer Dictionaries. (n.d.). General Inquirer Categories. Retrieved December 12, 2021, from http://www.wjh.harvard.edu/~inquirer/homecat.htm

Griffin, J. M. (n.d.). *Bitcoin*. Wikipedia. Retrieved December 12, 2021, from https://en.wikipedia.org/wiki/Bitcoin

Li, F. (2010). The information content of forward‐looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, *48*, 1049-1102.

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Accounting Research*.

McDonald, B., & Loughran, T. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. *The Journal of Finance*, *30*.

McGurk, Z., Nowak, A., & C. Hall, J. (2020). Stock Returns and Investor Sentiment: Textual Analysis and Social Media. *Journal of Economics and Finance*, *44*.

Resources // Software Repository for Accounting and Finance // University of Notre Dame. (n.d.). Software Repository for Accounting and Finance. Retrieved December 12, 2021, from https://sraf.nd.edu/textual-analysis/resources/

Tsay, R. S. (2005). *Analysis of financial time series*. John wiley & sons.