Dhirubhai Ambani Institute of Information and Communication Technology

Masters of Science in Data Science

Big Data Processing

Crypto currency Sentiment Analysis

Project Report

Sem II (AY 2022-2023)

By

Seat no.	Student
202218044	Shivani Pande
202218045	Swapnil Sheth
202218052	Vipul Singh
202218059	Jitul Bakshi

Project Guided by-

Dr. Sandip Modha

About the project

In recent years, cryptocurrencies have become a popular investment, attracting the attention of investors, researchers, and technology enthusiasts alike. Cryptocurrency is a digital or virtual currency that uses cryptography for security and operates independently of a central bank. Cryptocurrencies like Bitcoin, Ethereum, and numerous altcoins, are decentralized and operate on a distributed ledger called the blockchain. With the proliferation of digital currencies such as Bitcoin, Ethereum, and numerous altcoins, the cryptocurrency market has become a dynamic ecosystem that operates 24/7, transcending geographical boundaries and traditional financial institutions. As a result, understanding the factors that influence cryptocurrency prices has become a subject of great interest and importance.

The relationship between news sentiment, people sentiments and financial markets has long been recognized, as market participants often rely on news and media coverage to make informed decisions. The public sentiment expressed in news articles can significantly impact investor behavior, shaping market trends and influencing asset valuations. In the context of cryptocurrencies, where sentiment and speculation play substantial roles, analyzing news articles and correlating them with cryptocurrency prices can provide valuable insights into market dynamics

There have been many studies and analyses done to understand the relationship between the price of cryptocurrencies and the sentiment of news articles and people. Sentiment analysis is the process of analyzing text data to determine the emotional tone behind it. In the context of cryptocurrencies, sentiment analysis can be used to gauge public opinion about a particular cryptocurrency.

One study published in the Journal of Risk and Financial Management found that there is a positive relationship between the sentiment of public views and news articles about Bitcoin and its price. The study used sentiment analysis to analyse these views about Bitcoin from 2012 to 2018 and found that positive sentiment was associated with an increase in Bitcoin's price, while negative sentiment was associated with a decrease in price.

Another study published in the Journal of Behavioural and Experimental Finance analysed the sentiment of posts and articles on the BitcoinTalk forum and found that positive sentiment was associated with an increase in Bitcoin's price, while negative sentiment was associated with a decrease in price. Thus, the price of Crypto currency is very dependent on public and news sentiments, as it is not regulated by any single entity.

Note: Sentiment analysis can provide insights into the relationship between the sentiment of people and the price of cryptocurrencies. However, it is important to note that other factors such as market demand, adoption rate, and regulatory developments also play a very significant role in the price of cryptocurrencies.

Tools & Technologies

CCXT Library:

CCXT stands for "CryptoCurrency eXchange Trading Library." It is an open-source library that provides a unified API (Application Programming Interface) for accessing and trading on various cryptocurrency exchanges. CCXT is written in Python and supports more than 130 cryptocurrency exchanges, making it a popular choice among developers and traders in the cryptocurrency space. It allows you to access market data, manage orders, and execute trades across multiple exchanges using a consistent set of methods.

Text Blob:

TextBlob is a Python library for natural language processing (NLP) tasks. It provides a simple and intuitive API to perform various NLP operations, such as sentiment analysis, part-of-speech tagging, noun phrase extraction, language translation, and more. TextBlob is a simple library which supports complex analysis and operations on textual data. It has semantic labels that help with fine-grained analysis.

News API:

News API is a service that provides developers with a programmatic interface to access and retrieve news articles and headlines from various sources around the web. It offers a convenient way to fetch news content in a structured format, enabling developers to integrate news data into their applications, websites, or services. With News API, you can retrieve news articles based on specific search queries, filter by various parameters such as source, date, language, and category, and fetch the latest headlines from popular news sources.

Pyspark:

PySpark is the Python API for using Apache Spark, which is a parallel and distributed engine used to perform big data analytics. It enables you to perform real-time, large-scale data processing in a distributed environment using Python.

Workflow

For implementation of this project, we have used ELT pipeline (Extract-Load-Transform). The ELT pipeline is a data processing pipeline that enables the replication of data from a source system into a target system, such as a cloud data warehouse. It consists of three main steps:

Extraction: In the extraction step, data is copied from the source system. This involves fetching data from various sources, such as databases, APIs, or third-party applications. The goal is to gather the necessary data and prepare it for further processing.

Loading: During the loading step, the extracted data is loaded into the target system, which can be a data warehouse, data lake, or another storage platform. This process involves transferring the data from the source to the target, ensuring its integrity and consistency. The target system serves as a central repository for the data.

Transformation: Once the data is in the target system, organizations can perform various transformations as needed. These transformations involve manipulating the data to meet specific requirements, such as cleaning, structuring, aggregating, or enriching the data. The transformed data is then ready for analysis, reporting, or other downstream processes.

Overall, the ELT pipeline provides a flexible and efficient way to process and analyse large volumes of data. It enables organizations to replicate and transform data from various sources, ensuring it is readily available for analysis, study trends, gathering useful insights and decision-making purposes.

Data Source

To implement this project, we require two types of data:

1. The price trend of various cryptocurrencies

2. News Articles and Information regarding the sentiments of people towards cryptocurrencies.

Price Trend:

The trends of crypto currencies are highly dynamic and operate round the clock. Therefore, to accurately capture their dynamism, we require comprehensive data on the analysis of price trends. To acquire this data, we have utilized the 'ccxt' Library, which provides us with in-depth insights into the price trends of various crypto currencies.

'ccxt' library provides us more than 130+ crypto exchanges via their APIs or URLs, out of which we have used Binance's API as it is the most widely used crypto exchange platform, and it allows us to fetch hourly data easily.

```
ace
alpaca
ascendex
bequant
bigone
binance
binancecoinm
binanceus
binanceusdm
bit2c
bitbank
bitbay
bitbns
bitcoincom
bitfinex
bitfinex2
```

```
bitflyer
bitforex
bitget
bithumb
bitmart
bitmex
bitopro
bitpanda
bitrue
...
xt
yobit
zaif
zonda
```

After that we searched for different symbols that were available in Binance's API to find a common currency to compare crypto prices with. Here the cryptos of our interest are - "Bitcoin, Ethereum, DogeCoin, Solana".

```
BTC/USDT
ETH/USDT
HSR/BTC
OAX/ETH
DNT/ETH
MCO/ETH
ICN/ETH
MCO/BTC
WTC/BTC
WTC/ETH
LRC/BTC
LRC/ETH
QTUM/BTC
YOYOW/BTC
...
DOT/USD:DOT-230929
XRP/USD:XRP-230929
```

We found USDT most suitable for comparing the prices. So we have used the following pairs of symbols to fetch the prices - "BTC/USDT", "ETH/USDT", "DOGE/USDT", "SOL/USDT".

News Articles and Headlines:

To gather sentiment data, we will utilize news streams and employ a news API. This API will enable us to capture the articles on a specific day regarding a particular cryptocurrency. By leveraging the news API, we can collect and analyse news articles, headlines, and other textual information related to cryptocurrencies. This approach allows us to gain insights into the sentiment of individuals or the general public, helping us understand the common views about a cryptocurrency.

Extraction:

We have obtained raw data from both News API and the CCXT library. The extraction process involved authenticating our access to the API and establishing data streaming capabilities.

For the crypto currency prices, we conducted the extraction on a daily basis, capturing cryptocurrency trends at regular intervals of two-hour over a period of fifteen days from 16/04/2023 to 30/04/2023, for each of the crypto symbol pairs. (The restriction of 15 days is imposed by the exchange platform and API)

	timestamp	open	high	low	close	volume
0	2023-04-16 00:00:00	30295.10	30319.04	30120.00	30236.95	2199.07591
1	2023-04-16 02:00:00	30236.96	30284.78	30224.43	30249.37	1185.18424
2	2023-04-16 04:00:00	30249.36	30340.06	30245.82	30306.32	1587.43772
3	2023-04-16 06:00:00	30306.32	30399.40	30300.00	30344.97	1333.85689
4	2023-04-16 08:00:00	30344.97	30353.96	30254.99	30279.34	1882.36017

This is the data frame for BTC/USDT. We have made 3 other similar data frames for other symbol pairs.

For the news articles, we have retrieved 12 news articles from news API for each day, over a span of fifteen days from 16/04/2023 to 30/04/2023. The timestamp of 12 articles in a day is distributed equally over 24 hours. For each type of crypto currency, we have filtered the news with specific keywords to get good quality news like - Bitcoin, BTC, Doggy Coin, Rage of BTC/USDT, Elon's Coin, Ethereum-Tether bond etc.

10	2023-04-16T12:15:36Z	Bank Of America Reveals Surprise Crypto 'Key D...	https://www.forbes.com/sites/digital-assets/20...	A Bank of America report has named a surpris...
11	2023-04-16T11:45:25Z	Regulators Under Pressure From Banking To Cryp...	https://www.forbes.com/sites/lawrencewintermey...	While U.S. mid-tier banks and regulators are s...
12	2023-04-17T13:36:54Z	US stocks trade mixed as traders wait for next...	https://markets.businessinsider.com/news/stock...	Bank of America, Goldman Sachs, and Tesla are ...
13	2023-04-17T15:37:07Z	Dogecoin has surged 24% in a month, outperform...	https://markets.businessinsider.com/news/curre...	Dogecoin added to gains on Monday, spiking 6%....
14	2023-04-17T20:15:37Z	US stocks edge higher as traders get ready for...	https://markets.businessinsider.com/news/stock...	Goldman Sachs, Bank of America, Morgan Stanley...

This is the data frame for bitcoin news. We have made 3 other similar data frames for other crypto currency news as well.

Loading:

The collected streamed data from the CCXT library and News API is saved in CSV file format.

Sentimental Analysis: We have done sentimental analysis using the text blob library on the description column of the streamed news. This gives us a sentimental score of the news as 1 for positive, 0 for neutral and -1 for negative. This is done for each type of crypto news separately. We made a list of sentiment scores and merged it into the corresponding dataframes, and saved into the csv file.

Finally the csv files are loaded into the pyspark environment as distributed datasets (Resilient Distributed Datasets or RDDs). Now using pysparck we will be able to transform and manipulate the data to find the correlations between two types of data.

Transformation:

We have sorted the News data frames according to timestamp in the ascending order.

In the price data frames we have splitted the date and time separately from the timestamp column.

Finally the two types of RDDs - price and articles, were merged using date as the base column, for each type of crypto currency.

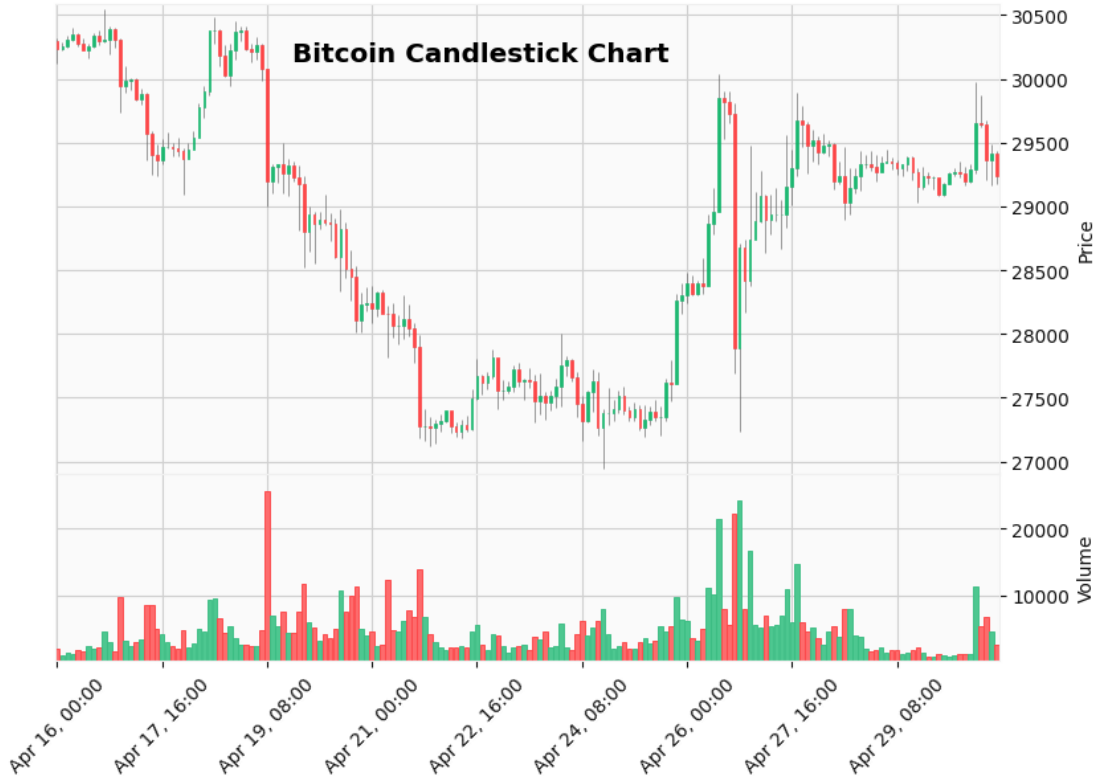
After that we have used closing price and sentimental score to derive relation between the crypto currency price and news article sentiments.

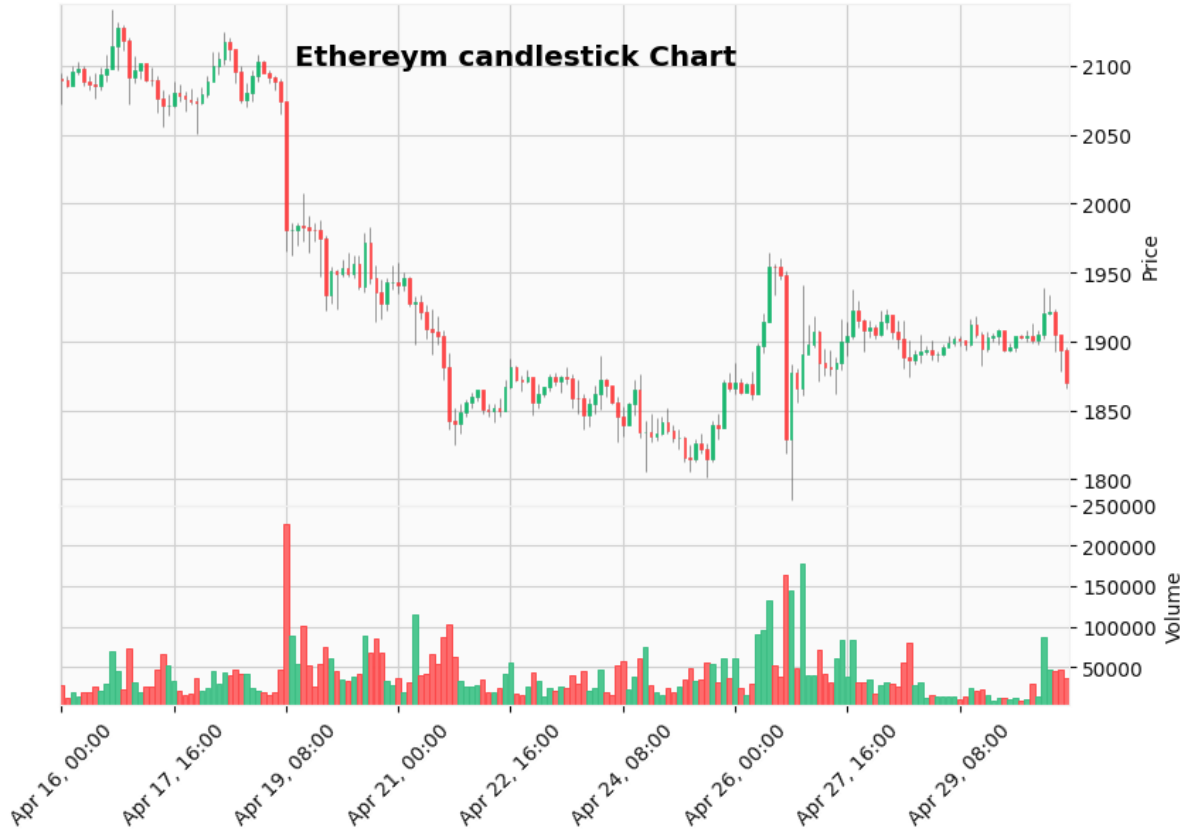
Final Goal

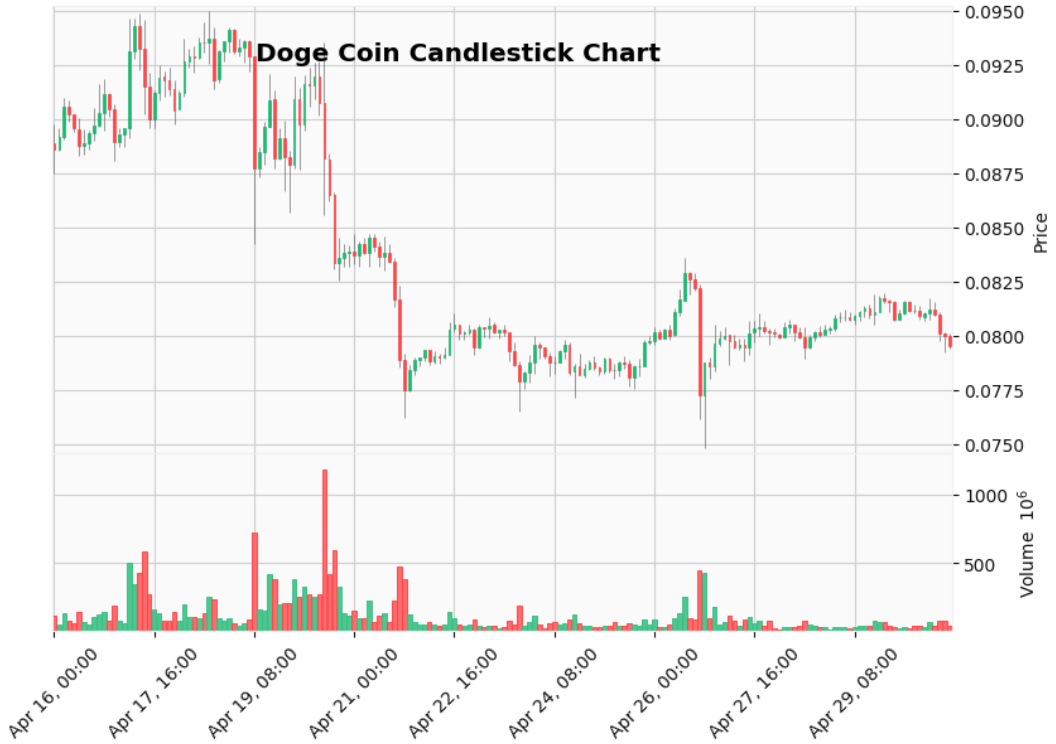
The final goal of our project, based on this data gathering and sentiment analysis, is to provide valuable insights to users. These insights can help users understand and assess the sentiment of the market towards cryptocurrencies. By having a clear understanding of sentiment, users can make more informed decisions when trading or investing in cryptocurrencies.

The project aims to contribute to enhanced profitability and risk management in the cryptocurrency market. By leveraging sentiment analysis, users can potentially identify sentiment-based trends, sentiment shifts, and market sentiments that may impact cryptocurrency prices. This knowledge can assist users in making more strategic decisions, potentially improving their profitability and reducing their exposure to risk

Result and Conclusion






Doge Coin Candlestick Chart
0.0950
0.0925
0.0900
0.0875
0.0850
0.0825
0.0800
0.0775
0.0750
Price
1000
500
Volume 10^6
Apr 16, 00:00
Apr 17, 16:00
Apr 19, 08:00
Apr 21, 00:00
Apr 22, 16:00
Apr 24, 08:00
Apr 26, 00:00
Apr 27, 16:00
Apr 29, 08:00


Solana Coin Candlestick Chart
26
25
24
23
22
21
Price
3
2
1
Volume 10^6
Apr 16, 00:00
Apr 17, 16:00
Apr 19, 08:00
Apr 21, 00:00
Apr 22, 16:00
Apr 24, 08:00
Apr 26, 00:00
Apr 27, 16:00
Apr 29, 08:00

Here we can conclude that there is no strong correlation between the news article sentiments and price trends of cryptocurrencies. We can see as the price of cryptocurrencies increases the sentiments are equally distributed among the three sentimental classes - positive, negative and neutral.

This correlation can be justified as -

When the price of a crypto currency increases, the bullish investor of that crypto currency will give out positive sentiments and the bearish investor will give out negative sentiments, while there will always be some with neutral sentiments.

Although this is not the desired correlation as given by actual facts. This may be due to not getting actual opinions of people via tweets and switching to news articles related to cryptocurrencies.

```
Correlation between sentimental value and closing price: -0.06062066184547464
```

```
Coefficients: [-93.18346223896233]
Intercept: 28821.27623185056
Correlation between sentimental value and closing price: -0.06496018116154084
```

References & Sources

- https://www.wikipedia.org/
- https://medium.com/
- https://towardsdatascience.com/beginners-guide-to-pyspark-bbe3b553b79f
- https://newsapi.org/
- https://docs.ccxt.com/