**Tweets related to the stock market - data analysis and visualization.**

**Group Name:**

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **First name** | **Last Name** | **Student number** |
| **Aanal** | **Patel** | **C0910376** |
| **Bimal Kumar** | **Shrestha** | **C0919385** |
| **Danilo** | **Diaz** | **C0889539** |
| **Ernie** | **Sumoso** | **C0881591** |
| **Jayachandhran** | **Saravanan** | **C0910392** |

**Submission date: 22 March 2024**

Contents

[Abstract 2](#_Toc162024722)

[Introduction 3](#_Toc162024723)

[Methods 3](#_Toc162024724)

[Data overview 3](#_Toc162024725)

[Validation and formatting 4](#_Toc162024726)

[EDA 5](#_Toc162024727)

[Text Processing 6](#_Toc162024728)

[Visualization 7](#_Toc162024729)

[Conclusions and Future Work 12](#_Toc162024730)

[References 12](#_Toc162024731)

# **Abstract**

This project compresses the analysis of the stock market-related tweets from popular social media X (also recognized as Twitter). The data source contains 8 separate CSV files mainly focusing on prices of the stocks of prominent organizations such as APPL, GOOG, and YHOO, including cryptocurrencies such as Bitcoins Altcoin, and Gold. The workflow of the project consists of data preparation and preprocessing, visualization, and data validating, after that, Exploratory Data Analysis (EDA) will be performed to gain the key insights of the data. The last step is to employ textual data processing techniques in order to analyze the text data of the datasets. After completing every step, some meaningful insights were uncovered, such as the rise in Altcoin-related tweets and the drop in tweets involving YHOO during the end of the third quarter. The day-wise tweet and the user counts are plotted.

**Keywords:** tweets, stock market, cryptocurrency, EDA, data visualization, text processing, data preprocessing.

# **Introduction**

In this project, various data handling techniques are applied to the given datasets, which consist of different capital or finance market details. The dataset provided is from the well-known social media platform X (formerly known as Twitter). Based on the given instruction, a complete analysis is carried out, and an extensive study of the data points is conducted. The following sections of the documents discuss in detail the applied methods and the results inferred from the processed data. The team split up the task into small teams which is discussed below in the table.

|  |  |
| --- | --- |
| **Team** | **Task Description** |
| Data Team | Responsible for understanding the dataset given and making it ready for pre-processing |
| Processing Team | Apply various processing and analysis techniques on top of the given data |
| Visualization Team | Generate graphical charts for a greater representation of the data points |
| Validate and Documentation Team | Validating the whole procedure and applying modifications. Document the whole process |

As per the above table, the team proceeded with the activity by clearly understanding the given scope. Later, the knowledge gained in respective fields is shared among the team members to distribute the learning outcomes equally.

# **Methods**

## Data overview

A total of 8 datasets is provided in the form of CSV (Comma Separated Value) file format containing tweet details of the major stocks like “APPL, GOOG, YHOO” and cryptocurrencies “Bitcoin, Altcoin, Cryptocurrency, CoinDesk” and finally, the tweet includes the commodity “Gold.” Each dataset contains 5 fields, which are explained clearly in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Field/Feature** | **Description** | **Datatype** |
| 1 | DateTime | Date and time of the tweet | Object → data time |
| 2 | TweetID | Unique ID for tweet on twitter | Numeric (int64) |
| 3 | Text | Tweet about the particular financial market | String |
| 4 | URL | Tweet URL (Link to access) | Object →String |
| 5 | User | User account URL (userId) | Object →Strong |

After thoroughly examining the given datasets, the distribution of each dataset is very similar and contains the same fields. The datatypes and the other aspects are discussed in the following section.

A computer code with text

Description automatically generated

*Figure:*

## Validation and formatting

Since all 8 datasets have common columns among themselves, steps to combining them as a single large dataset are taken care of, the initial step of validating the datatypes of data present in the various datasets is verified by comparing with all 8 datasets. The column names are standardized, and the final check of data is completed before concatenating the datasets. The reason for concatenating the dataset is to bring the distinct values of data with the same column names under one data frame for easy and smooth processing. It also allows a hassle and error-free approach to applying the visualization in the following steps.

A new column named “Tag” is created, which represents the financial topic of the dataset from which it came. A total of 8 tags are mapped with the combined rows, making the columns count to six.

A screenshot of a computer

Description automatically generated

*Figure:*

## EDA

Before proceeding with the processing method, an extensive Exploratory Data analysis is carried out on the combined dataset. The final data frame contains 36831 records/rows with 6 distinct feature sets present in it. The contribution towards the final dataset is described clearly in the table below.

|  |  |
| --- | --- |
| APPL | 3181 (8.64%) |
| ALTCOIN | 5001 (13.58%) |
| GOOG | 5001 (13.58%) |
| BITCOIN | 5001 (13.58%) |
| CRYPTOCURRENCY | 5001 (13.58%) |
| GOLD | 5001 (13.58%) |
| YHOO | 3644 (9.89%) |
| COINDESK | 5001 (13.58%) |

A screenshot of a computer

Description automatically generatedThe combined data is checked for general parameters like Missing, Duplicates and Unique Values present in the data points. In addition, various other data analysis is carried out to verify the data integrity. Surprisingly, the combined dataset has no missing and 1.25% of duplicate values, i.e., 460 rows. This provides us with a clear observation that the given dataset is properly cleaned before assigning.

* The Tweet ID and URL show 99% of unique values present in the whole 38K data.
* It is followed by Text(tweet) and date time with unique values of 97% and 79%.
* From the given dataset, only 50% of users are identified as unique, which indicates the user has tweeted more than one financial market-related tweet.
* The Duplicate values are removed as the impact is much less compared to the original dataset length.

*Figure:*

In further analysis in the future, the text length and various date-time parameters like day, month, and hours can be added for granular level detail.

## Text Processing

In the provided data, more than 50% are in text (string) format. Particularly, the processing of tweets requires the “text” column to be processed thoroughly. This text data is considered the most valuable data present in the entire dataset as it adds more value to the analysis. The text data contains numerous noises, so the initial step is to import the NLTK module to handle the text data.

* The tweet contains alphabets, numbers, white spaces, emojis and non-alphabetic words.
* The words of length no more than one are removed as the weightage contributed to the text is little.
* Standard text processing steps like punctuation removal, number extraction and keeping only alpha values are carried out.
* Social media is known for its instant method of sharing any news using artistic representation like emojis or GIFs. In text data, emojis are the major means of conveying messages to viewers.
* Emojis are removed, as they require heavy computational efforts to process and analyze.
* Non-alpha values are removed from the text.

In future processing, the ASCII method to handle non-English words or multi-language processing shall be applied. Along with this, various combinations like lowering the text, removing stop words and other NLP-based processing techniques can be applied.

## Visualization

The processed data is now ready for plotting against good graphical representation to gain more insights, and it is also easy for common people to interpret the results clearly without the need for analytical skills. In this project, various visualization methods are applied, including word cloud, line chart, and comparison chart.

**Word cloud:**

* Using word cloud, the text data is represented with the most common words being presented with large text size and the least common one in smaller size.
* It gives a clear understating of the word present in the text and the famous word listed in the data or discussion.
* A collage of words

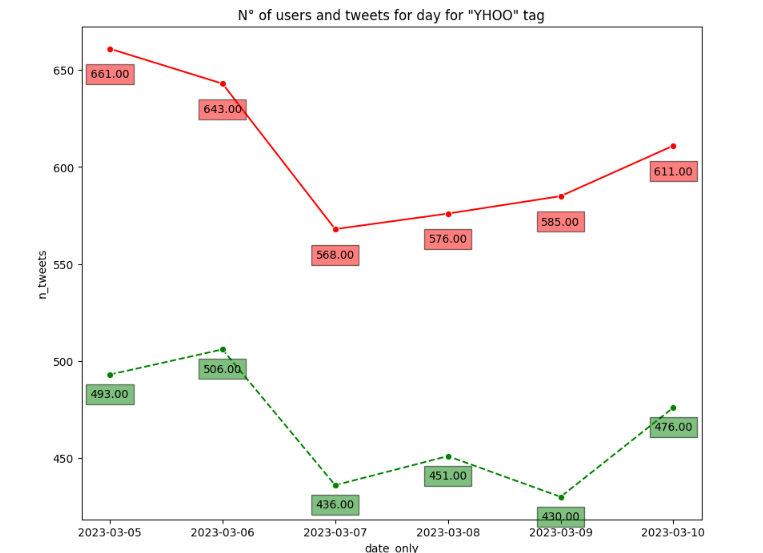
  Description automatically generatedTable representing the most common words and least common words present in each tag are discussed.
* The image represents the word cloud of APPL tag, where the words are scattered around with varying text sizes.

*Figure:*

|  |  |  |
| --- | --- | --- |
| **Tag** | **Most Common Words** | **Least Common** |
| APPL | Google, Tesla, NASDAQ | ABIO, UAVS |
| ALTCOIN | PAW, TREATS, BOT | BTC, time |
| GOOG | AMZN, ANALYST, FB | GME, LEN GE |
| BITCOIN | BTC, SVB, MONEY | block, USDC |
| CRYPTO | - | UTC, silver gate |
| GOLD | BETTER, PEOPLE, YEAR | hope, show |
| YHOO | BBT, now, KHOSI | today, must, hayi |
| COINDESK | COINDESK | Por, RT |

The above table gives a clear understanding of the words utilized most in each tag set and helps to a greater extent to analyze the topics discussed in each tweet.

**Tweet and User Counts:**

The number of tweets for each category is plotted and the count of users present in it is also calculated. It is noticed that for most of the stock market groups, the number of tweets and users keeps growing over time, except for the YAHOO tag, which started decreasing around May 2023 and then slowly came back until October 2023.

Tags like bitcoin, cryptocurrency, and Gold; there are single points data, meaning that most of these data come from a single day. The daily number of tweets per keyword and per user are plotted for 8 different stock groups, the datetime column is considered to be in the correct format, and loops

A screenshot of a computer

Description automatically generated*Figure:*

through the tags and plot the number of tweets and users over time using a line plot. The red line represents the number of tweets, and the green line represents the number of users.

*Figure:*

A graph with numbers and lines

Description automatically generated with medium confidence

*Figure:*

* Most of the categories follow a similar behavior on the number of users.
* **A graph with numbers and lines

  Description automatically generated**‘APPL’ has the most number of data points (6) resulting in a standard increase of tweets whereas the ‘’GOOG” started with exponential rise but dipped in the final.
* “Bitcoin” and “Crypto” tags represent the highest number of tweets present totaling 10,002 (5001,5001) tweets per day
* “Gold” has more number of users tweeted about in a particular day.

*Figure:*

* A graph with red and green lines

  Description automatically generatedHowever, 'altcoin' presents an unusual spike just at the right end of the timeline (September - October 2023).
* All categories have slowly increased their number of users over time.

*Figure:*

* The number of counts and the user plot gives a clear view of public interest and familiarity. Further, this can be used to generate various prediction-based model and statistical analyses on the current trends.
* The day-wise or month-wise plot signifies the impact of the price and growth of certain stock groups.

**Inference**

* 'Altcoin' presents an unusual spike of over 4 thousand unique tweets from September - October 2023. The ‘YHOO' category has instead decreased the number of unique tweets over time.
* Overall, it uncovered a hidden pattern within the tweets, demystifying the growth of tweets and users for the 2023 year for each category and identifying unique outlier behaviors (altcoin and Yahoo).
* Unique tweets over time also display a slow growth over time for the year 2023.
* This complete study gave us more understanding of the tweets related to the crypto markets and the commodities. The distribution of tweets across various timelines is plotted to learn the hidden pattern.

# **Conclusions and Future Work**

Thus the complete study of the given dataset is concluded and the results are discussed in detail with the respective charts as visual representation. It consists of steps starting from data pre-processing to NLP-based text analysis. The future scope of the project includes.

1. Name entity-based advance NLP processing techniques on the text data.
2. Day and time-wise tweet analysis to get the granular level analysis of data.

In the future, by using advanced Natural Language Processing (NLP) we can gain deeper insights of public opinion on their stock prices by extracting the sentiments of the tweets. It can also help in creating machine learning models that can predict stock prices and market trends.

# **References**

[1] <https://medium.com/analytics-vidhya/how-to-use-google-colab-with-github-via-google-drive-68efb23a42d>

[2] <https://pandas.pydata.org/docs/reference/api/pandas.concat.html>

[3] <https://usharbudha-dev09.medium.com/eda-of-stock-market-using-time-series-9662fd18bfc5>

[4] <https://towardsdatascience.com/finding-temporal-patterns-in-twitter-posts-exploratory-data-analysis-with-python-8aac618c8699>