Stock Market Tweet Analysis

Parvathy Vysakh
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0818539@mylambton.ca

Patricia Adolph
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0816792@mylambton.ca

Pooja Selby
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0821687@mylambton.ca

Suchithra Chandrasekharan
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0816811@mylambton.ca

***Abstract*— In the past decade, Twitter has experienced a tremendous growth worldwide in line with the success of social networks. An online Application Programming Interface (API) is made available for any developer willing to use Twitter data. Moreover, tweets usually contain hashtags and symbols that facilitate the search for relevant posts. Stock data have several interesting values to investigate. In the following study, we analyzed the collected tweets related to stock market keywords. The data for a period is collected and studied with respect to the frequency of the keyword usage and the number of people posting tweets regarding the same.**

***Keywords*— Altcoin, Bitcoin, Coindesk, Cryptocurrency, Gold, APPL, GOOG, YHOO**

I. Introduction

Twitter, a social networking site launched in 2006, is undoubtedly one of the most popular social media platforms available today, with 100 million daily active users and 500 million tweets sent daily. The Twitter platform helps us in identifying the latest trending topics. By analyzing the number of retweets / re-shares of the tweets in Twitter we can understand the user's interests. In our examination, we will be focusing on analyzing the stock market trends by collecting information on chosen catchphrases.

II. Tickers under study

In this analysis, we fundamentally concentrate on eight catchphrases which incorporates generally of stock showcase tickers/symbols speaking to driving stocks within the current economy. The tweets containing these keywords will be extracted for the study. The selected keywords are as follows,

A. Cryptocurrency

Cryptocurrency is a form of payment that can be exchanged online for goods and services. Many companies have issued their own currencies, often called tokens, and these can be traded specifically for the good or service that the company provides. Cryptocurrencies work using a technology called blockchain [6].

B. Bitcoin

Bitcoin is known as a type of cryptocurrency because it uses cryptography to keep it secure. There are no physical bitcoins, only balances kept on a public ledger that everyone has transparent access to (although each record is encrypted). All Bitcoin transactions are verified by a massive amount of computing power via a process known as "mining" [7].

C. Altcoin

Altcoin is a cryptocurrency alternative to Bitcoin — its name is a portmanteau of "alternative" and "coin." Since Bitcoin is widely regarded as the first of its kind, new cryptocurrencies developed after are viewed as alternative coins — or altcoins [8].

D. Coindesk

CoinDesk is a media outlet that strives for the highest journalistic standards and abides by a strict set of editorial policies. CoinDesk is an independent operating subsidiary of Digital Currency Group, which invests in cryptocurrencies and blockchain startups [9].

E. GOLD

Gold stocks are those of publicly traded companies and exchange-traded funds (ETFs) that are focused on gold. The industry consists of these types of entities: Mining companies: These are the companies that mine and sell wholesale gold [10].

F. APPL

APPL is a hashtag used to represent the stock updates of the Apple company. Apple is one of the world's largest technology company with respect to its revenue. Apple Inc. is a global technology company that designs, manufactures, and sells smartphones, personal computers, tablets, wearables, and accessories [11].

G. GOOG

GOOG is a stock ticker for the stocks of ALPHABET, the Google's parent company. Apart from GOOG, the company has another class of stocks, called GOOGL. The main difference between these two classes of stocks is that the former does not have any voting rights [12].

H. YHOO

YHOO is the symbol used to represent the stocks of Yahoo Inc. Yahoo is a web service introduced in the year 1994. The company went public in the year 2000 [13].

III. Twitter API with Python

The Twitter API enables programmatic access to Twitter in unique and advanced ways. We can use the *Twitter API* to retrieve recent tweets from users and retrieve tweets with certain hashtags. The Twitter API (Application Programming Interface) enables developers to extract and write contents on Twitter. The Twitter APIs can only be accessed if the Twitter has granted the user with developer access. The developer access gives the user special privileges and enables the user to make use of Twitter APIs.

In this project, we first applied for a developer account and had our use case approved. In order to access all the required API Keys and authorization credentials it is imperative that we create a Project and App to find or generate the API credentials to access Twitter.

To access the Twitter API using Python, we use an open-source Python package called Tweepy. This package includes all the required functions to handle the Twitter API and helps the user with the authentication and http requests [2].

The Twitter API authenticates every request with OAuth; therefore, the user should generate the required credentials. The Twitter's OAuth authentication requires four credentials, they are as follows:

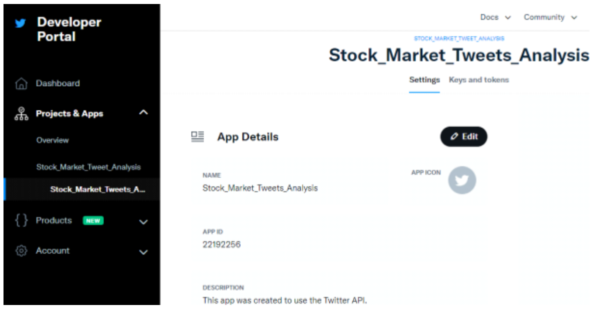
A. *Consumer Key* - This is essentially a username, and allows you to make a request on behalf of your App.

B. *Consumer Secret* - This is a password, and allows you to make a request on behalf of your App.

C. *Access Token* - This token represents the Twitter account that owns the App, and allows you to make a request on behalf of that Twitter account.

D. *Access Secret* - This token also represents the Twitter account that owns the App, and allows you to make a request on behalf of that Twitter account.

These four credentials can be generated from the Twitter dashboard. To access the functionalities and generate the credentials, we will first define an app in under our twitter profile. Once the credentials are generated from the Twitter app settings in the Keys and Access Tokens tab, we can start our analysis by collecting the required data from the Twitter database.



Stock_Market_Tweets_Analysis

NAME
Stock_Market_Tweets_Analysis

APP ID
22192256

DESCRIPTION
This app was created to use the Twitter API.

Fig. 1. Twitter Developer Portal.

IV. Data Extraction, Cleaning & Visualization

A. *Data Extraction*

Tweepy is a Python library for accessing the Twitter API. It is great for simple automation and creating twitter bots. For any analysis, the first and foremost step to initiate the study is by gathering the required data. In our project, we will be collecting the tweets related to our subjects of study. We will make use of the Tweepy library and Twitter API to extract the required data from the Twitter database.

To collect tweets in real time is the very first step for two purposes:

- Create the dataset for the ML model training purpose.
- The streaming will be used to demonstrate the real-time analysis.

In this analysis, we will be concentrating on the tweets posted for a period of one week and save the collected tweets into csv file based on each stock related keyword. As the first step towards the process, we created the authentication object using the authentication credentials in a dictionary format and also set the access token. The twitter_auth() is the user-defined function that has the authentication credentials to establish connection between twitter and python using the tweepy library. Thousands of tweets are shared on Twitter per second, hundreds of millions per day, hundreds of billions per year – all of which is an absolutely untapped resource for gathering data. Stock data have several interesting values to investigate and our project is focused on 8 such ticker symbols. The ticker_symbols is the list of 8 stock related keywords namely 'Altcoin', 'Bitcoin', 'Coindesk', 'Cryptocurrency', 'Gold', 'APPL', 'GOOG' and 'YHOO'. Furthermore, start_date and end_date variables indicate the starting date and ending date of tweet collection (13th to 19th October 2021). The ticker_symbols, start_date and end_date are the arguments passed in the user-defined function get_tweets_csv(). The User-defined function get_tweets_csv() will collect the tweets related to the keywords in 'ticker_symbols' for one week and will save the tweets for each ticker_symbol as separate csv files based on the 'ticker_symbol'.

We created the API object to get tweets using the authentication object and collected tweets for a week for each ticker. A cursor object is created using 'tweepy.Cursor'. The class constructor receives the aforementioned API method to use as the source for results. The Cursor object has an items() method that returns an iterable you can use to iterate over the results. Here we intend to collect 500,000 tweets for each ticker symbols for the one week duration. The 'extended_tweet' object provides the 'full_text' field that contains the complete, untruncated tweet message when longer than 140 characters. The dictionary tweets_stock_dict combines all the tweets of each ticker and convert it into a data frame and ultimately saved as a csv file for each ticker. We have collected a total of 2161018 tweets from twitter related to the stock market tickers for twitter analysis.

TABLE I. TOTAL NUMBER OF TWEETS PER TICKER

Ticker_Name	*No. of Tweets*
AltCoin	124400
BitCoin	700001
CoinDesk	31951
CryptoCurrency	580384
Gold	700001
APPL	1742
GOOG	5306
YHOO	17241

B. *Data Cleaning*

Data cleansing is an essential part of data science. Working with impure data can lead to many difficulties hence it requires fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset . By cleaning data it removes major errors and inconsistencies that are inevitable when multiple sources of data are getting pulled into one dataset.

For further preprocessing, we will be using glob (global) package which is a useful part of the Python standard library to retrieve all file paths that match a specific pattern. The glob.glob loops through destination folder to return all the .csv files. The pandas library will be reading the matched files' content into a new dataframe and we will be adding a column for the corresponding stock keywords for each tweets.

The extracted twitter data consist of punctuations, duplicates, numbers and words with length lesser than 2 characters that fail to add much value to the further analysis. The drop_duplicates() function in the pandas library will be used to remove the duplicates from the dataframe whereas the punctuations, numbers in tweets and short words can be eliminated using the replace method which takes in the required regular expression as input parameter. The data cleaning phase detected a total of 82 duplicate tweets which will be removed by the drop_duplicates() function. This ensures that only unique tweets will be considered for the data visualization phase.

C. *Data Visualization*

Visualization provides a deeper insight to larger datasets and helps us in identifying patterns among various data. In our project, we will be analyzing the obtained dataset through various visualizing tools.

In order to present the daily number of tweets as well as the daily number of users, for each keyword, we will be using the pandas library groupby() function. The groupby() initially groups the ticker and tweet_date to get the total number of tweets per day. Consequently, the total number of users and the total number of tweets for each ticker on a daily basis is also filtered out. The merge function (pandas library) coalesces the 'ticker', 'tweet_date' and 'user_id' columns into a single dataframe, 'ticker_grp' that will be the input for visualization.

For the graphical representation of each tickers, we will be creating a static grouped bar chart with multiple (double) bars with the help of Python libraries: Pandas, Matplotlib, and Seaborn. We plotted and visualized the graph for each ticker on a daily basis for a week and will be saving each plot as a 'png file' with the ticker name as the filename.

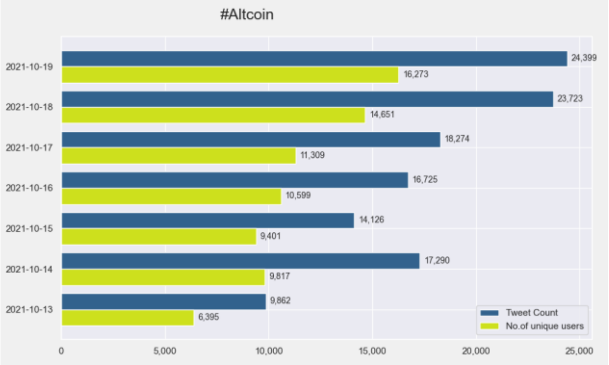


Fig. 2. Visualization of tweet counts and number of users of Altcoin.

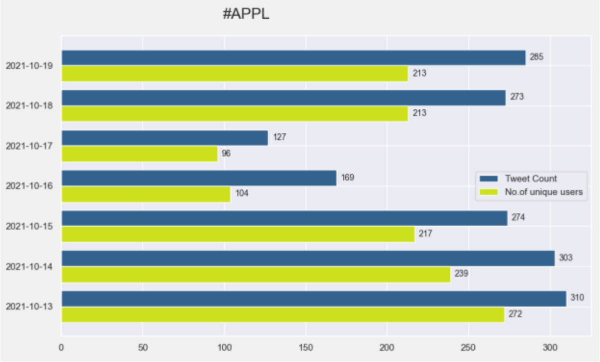


Fig. 3. Visualization of tweet counts and number of users of APPL.

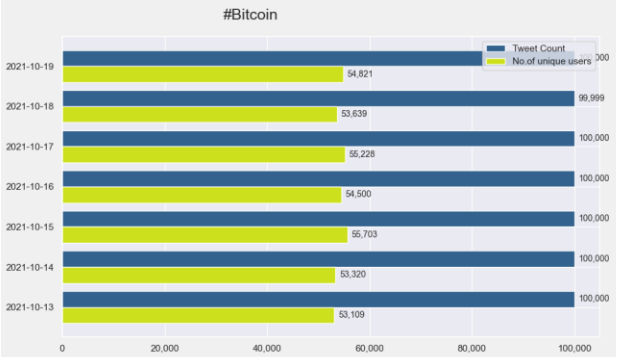


Fig. 4. Visualization of tweet counts and number of users of Bitcoin.

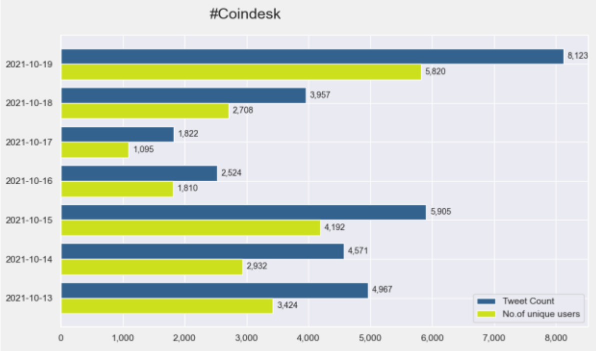


Fig. 5. Visualization of tweet counts and number of users of Coindesk.

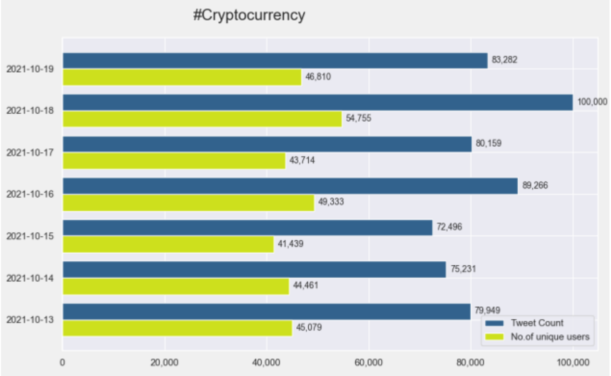


Fig. 6. Visualization of tweet counts and number of users of Cryptocurrency.

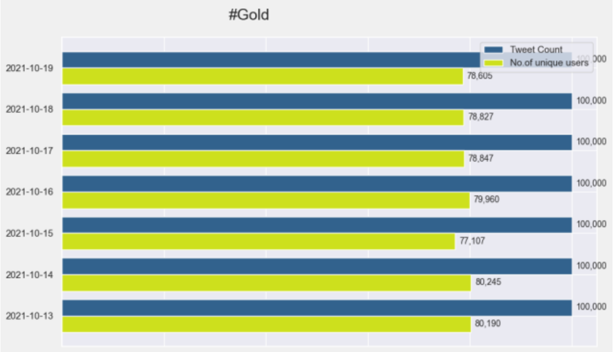


Fig. 7. Visualization of tweet counts and number of users of Gold.

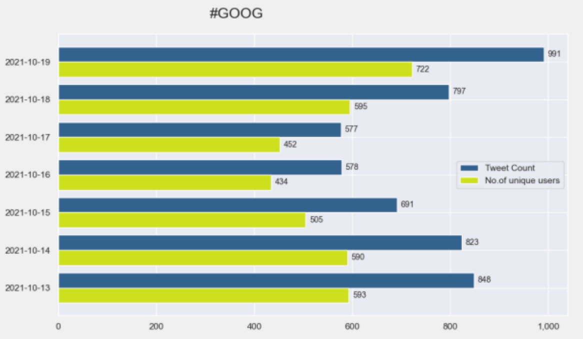


Fig. 8. Visualization of tweet counts and number of users of GOOG.

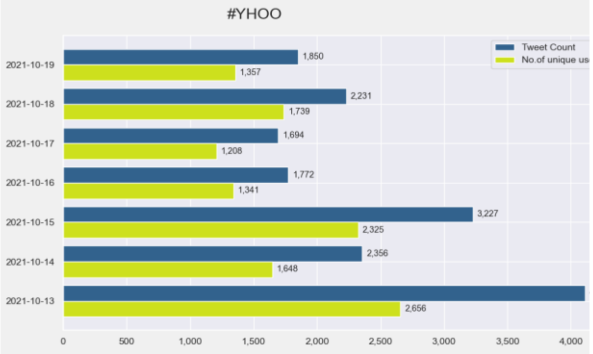


Fig. 9. Visualization of tweet counts and number of users of YHOO.

From the above visualization, it is evident that Cryptocurrency, Gold and Bitcoin are trending for the week of the analysis based on the number of tweets and number of users.

V. Conclusion

Evaluating the impact of tweets on stock market is a major challenge and has been addressed several times for the indices. The result of our twitter analysis through visualization proves that the top trending stocks in stock market taken between the duration of a week from twitter are Bitcoin, Gold and Crypto-Currency whereas the least trending stock in stock market is APPL. We explained how we examined, extracted and preprocessed the data corresponding to the tickers under examination. Eventually, by visualization tools in python we build a graphical representation of stock market movements with various Twitter variables such as tweet counts, tweet date and user.

References

[1] J. Bollen, H. Mao, X. Zeng "Twitter mood predicts the stock market" J Comput Sci, 2 (1) (2011), pp. 1-8

[2] E. Bakshy, J.M. Hofman, J.D. Watts, W.A. Mason "Everyone's an influencer: quantifying influence on twitter"

[3] J. Berger, K.L. Milkman "What makes online content viral?" J Market Res, 49 (2) (2012), pp. 192-205

[4] Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259-1294.

[5] Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1310-1319). Association for Computational Linguistics.

[6] Bollen, J., Counts, S. & Mao, H. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051.

[7] Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 450-453.

[8] Brown, E. D. (2012). Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. Proc. of SAIS.

[9] Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294.