Predicting the Market Movement of Ethereum Using Twitter Sentiment

Prince Jan Kumi[1]

[1] University of Ghana, Legon, Ghana
pdgyan@st.ug.edu.gh

Abstract. As with any other industry, cryptocurrency and its market are also influenced by a variety of factors. These can range from general sentiments about blockchain technology and digital currencies, to more specific events like exchange hacks or partnerships between different companies. Twitter has developed into a huge social tool used all over the world for the exchange of ideas and views. As a consequence of this, cryptocurrency investors and traders are looking to utilize the platform so that it can assist them in making sound financial choices. It has been demonstrated that the sentiments gleaned from the contents of Twitter, specifically tweets and comments, are a reliable indicator of the movements of markets. The objective of this paper is to investigate and further explore how sentiments and volume of data from twitter can influence the movement of Ethereum Market. A number of different classification models were developed based on three features; for a given day, an average sentiment score of tweets, the volume of tweet and an interest score from google trends. Random Forest Classifier and Naïve Bayes model were trained to predict the change in Ethereum market given those features.

Keywords: Cryptocurrencies, Ethereum, Twitter, Sentiment Analysis, Market Change, Random Forest algorithm, VADER, Naïve Bayes, Support Vector Machines.

1 Introduction

1.1 Background

Cryptocurrencies have been making waves as of late. Even if you've been living under a digital rock, chances are you've heard about Bitcoin, the first and still the most popular cryptocurrency. A cryptocurrency is a digital asset that is stored on an individual or company's digital wallet and can be used as a medium of exchange for goods, services, and other virtual currencies. Cryptocurrency can also be described as a type of digital token — a string of code — that you can buy, sell, and trade. It's also a type of commodity like gold or oil. The primary difference between cryptocurrency and other commodities is that it's not just a digital representation of an existing asset (like gold). Instead, it's its own standalone asset with its own market value.

Twitter is the is the 9th most visited website globally(Hootsuite, 2022) and the third most popular social network among cryptocurrency investors. A recent survey found that nearly a quarter of active crypto users are on Twitter, making it the third most popular social media channel for this target group after Instagram and Reddit.
Additionally, one stand-out player in the world of cryptocurrency is Ethereum, a public blockchain network similar to Bitcoin, but with additional features that make it more suitable for business applications. It has the second largest market capitalization. Specifically, Ethereum enables the creation of smart contracts, which are self-executing computer programs that can be used to implement contract logic and monitor obligations between parties. Ethereum also has its own native currency (or ICO token), called Ether. However, unlike Bitcoin where the primary use case is as a fungible digital store of value, Ether has a variety of use cases on the Ethereum network itself.

1.2 Problem

The cryptocurrency market is prone to consistent fluctuations, just like any other market out there. In order to be a knowledgeable investor, it is necessary to keep up with all of the news that pertains to the market. This includes keeping track of everything from sudden dips in price to unexpected increases, as well as everything in between. The same is true for prospective traders who could be interested in joining the market but are unsure about the appropriate time to act.

1.3 Objective

This paper presents a solution for locating patterns in the Ethereum cryptocurrency market and forecasting changes in the Ethereum cryptocurrency market based on the information provided. Given the volume of tweets mentioning or relating to the Ethereum coin for the day, the average sentiment of tweets relating to the Ethereum coin is taken into consideration. Traders and investors who are just getting their feet wet in the bitcoin market might stand to gain by utilizing this strategy.

2 Literature Review

2.1 Related Work

Because of the large availability and accessibility of data pertaining to these markets, it has been shown that cryptocurrency markets can be a topic with a great deal of potential for research in financial time series issues(Valencia et al., 2019). This study builds on the work of other academics in the behavioral sciences and design science, who have explored a wide variety of issues and concepts in their respective fields. To this day, a significant amount of research has been carried out on sentiment analysis, which has resulted in the development of a variety of methods that can detect the presence and magnitude of opinions expressed by users in unstructured text, such as user-generated reviews or social media posts.

There is no question that public opinion may have an effect on the price of a good or an asset. Although this statement is referring to a body of research that has a solid foundation and investigates the application of sentiment analysis to traditional markets(Mittal & Goel, n.d.; Rao & Srivastava, 2012), sentiment analysis can also be used for predicting the price of cryptocurrencies, as has been demonstrated in a substantial amount of recent work. One major work done in the paper "Bitcoin price change and trend prediction through twitter sentiment and data volume" (Critien et al., n.d.) achieved an accuracy of 63% in predicting the magnitude of change (Increase/Decrease) in prices. While various works have been done on sentiment analysis for predicting market movements, most focus on the stock market and Bitcoin when it comes to cryptocurrency. This project seeks to draw inspiration from such existing works and apply them to the Ethereum cryptocurrency.

3 Data

Data used in this research was collected from various sources all spanning 153 days from 1st March 2022 to 31st July 2022. In addition to the sentiment as an input feature, a few other inputs were considered in order to solve the problem of price changes. The second was tweet count where the number of tweets on Ethereum coin was considered. The third was the interest over time in the Ethereum Cryptocurrency.

3.1 Tweets from The Twitter API

Due to the need for historical data from twitter, I began by implementing helper functions that connect to twitter's full archive search endpoint utilizing the python requests library. Before the average sentiment for a day could be calculated, at most a 1,500 tweets per day for 153 days were pulled from the twitter API. The Twitter API allows for filtering of tweets, for this reason, the first step for retrieving the tweets was the need to create a search query keyword like ("Ethereum"). In addition to that, the language of tweets needed was also specified as such ("Ethereum lang:en").

The following step was to provide the start time and finish time for the tweets that were required and then to send those times via the helper functions so that a connection could be made to the endpoint. After iterating over the dates that were between the start and finish times that were given, a maximum of 1500 tweets per day were gathered and stored in a csv file.

3.2 Collecting Tweet Counts

The twitter API also provided an endpoint for collecting the total number of tweets relating to a specific keyword or topic. By specifying the start and end dates for a helper function, tweets volume was collected for each day from 1st March 2022 to 31st July

2022 and saved in a pandas Data Frame and further in a csv file. Fig. 1 shows the distribution of tweets count.

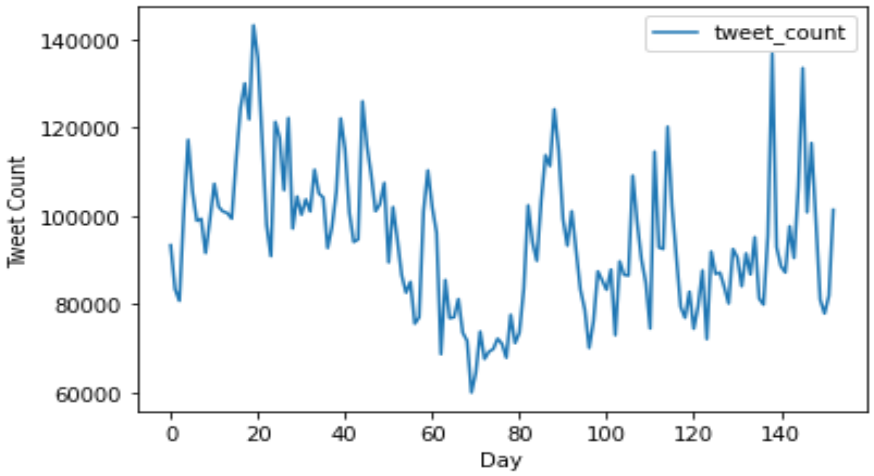

Fig. 1. A line chart showing the distribution of tweet count across the hundred and fifty-three (153) days.

3.3 Collecting Historical Data on Ethereum Prices.

Dealing with market changes means dealing with changes in Ethereum prices. Various services that provide access to historical hourly and daily prices exist, but this project utilized Yahoo Finance, a media property of Yahoo! Network. Stock quotations, press announcements and financial reports are just a few examples of the financial news, data, and analysis that may be found on the Yahoo! Finance media domain. The Ethereum prices from March to July 2022 was retrieved manually (not automated with code) from their website and saved into a csv file.

In addition, in order to forecast how the market would move, whether it would go up or down, the open price and the final price of each trading day were calculated and included. It was determined by taking the difference in price between the open and the closure of the market, dividing that number by the opening price, and then multiplying that number by one hundred.

$$\text{Percentage Change} = \frac{(Closing\ Price - Opening\ Price)}{Opening\ Price}\ x\ 100 \quad (1)$$

After the change was calculated, the positive or incremental changes were labeled as 1 and the negative or the decremental changes were labeled as 0. As a result, a new categorical feature, “Change” was added to the main dataset. It was either a 1 or 0 where 1 means a positive magnitude of change and 0 means a negative magnitude of change.

3.4 Collecting Trends from Google.

Google Trends is a platform that provides access to insights and information on topics and search terms on Google. It provides access to data as far back as 2004. It scores the interest over time of a particular topic or search term either by regions or by related topics. Interest over time is scaled between 0 and 100. The global interest over time in "Ethereum" from March to July was collected manually from the website and saved in a csv file.

3.5 Cleaning Tweets

Tweets collected from the twitter API comes with a lot of noise such as hyperlinks, hashtags ("#"), handle symbols ("@"). As a result, further processing is required to remove all characters that do not provide any information for sentiment analysis.
The first processing a tweet goes through is the removal of handles ("@") and hyper-links using a regular expression to substitute phrases that match the regex expression with an empty string. A regular expression such as "`http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+`" matches expressions like `https://t.co/l4jcYke7rw` and (@[A-Za-z0-9_]+) matches expressions such as @dev. Links such as those above are not relevant and do not contribute anything useful to the sentiment analysis process. The same as twitter handles, which are used to identify certain usernames on twitter.

Additionally, tweets are further processed by tokenizing each tweet by word and removing punctuations, odd characters like # $% ^& and single characters. After tokens were filtered, they were joined back forming a complete phrase or sentence, good enough to be used for the sentiment analysis.

Table 1. A table comparing data before and after it is cleaned.

Before Cleaning Data	After Cleaning Data
RT @wragstoriches: Years of being called slow and here we are with #Cardano and #Ethereum at the same general level of core development.	rt years of being called slow and here we are with Cardano and Ethereum at the same general level of core development

3.6 Normalizing data

The final dataset had its values in the columns covering different ranges. Tweet count ranged from 6000 – 10000, trend ranged from 0 – 100, change was either a 1 or 0. Different ranges affect the performance of our model, consequently the dataset was scaled using the MinMaxScaler from the scikit learn library

4 Analysis

4.1 Numerical Exploratory Data Analysis

Numerical exploratory data analysis was performed using various functions from the Pandas library in python. The data type of the **sentiment, count, trend and change** were all floats (64 bits). Moreover, there were hundred and fifty-three (153) data values for all the features with no missing values.

4.2 Sentiment Analysis

In this research, a pretrained sentiment model was used to classify the sentiment of tweets. This lexicon and rule-based sentiment analysis tool is called VADER (Valence Aware Dictionary and sEntiment Reasoner) and it is especially geared to the emotions that are expressed in social media. It is completely free and open source.
An integral and pretrained sentiment analyzer known as VADER is already included in NLTK (Valence Aware Dictionary and sEntiment Reasoner). Because VADER does not require training, getting results from it is much faster than with many other types of analyzers. However, VADER works best with the type of language that is typical of social networking platforms, such as brief phrases peppered with slang and acronyms. When evaluating lengthier, more organized sentences, its accuracy decreases, but it's still a reasonable starting point in many cases. VADER analyzer does not just classify phrases as just negative or positive, but gives a polarity score containing negative, neutral, positive, and compound score. The compound score, though not the average, encapsulates the overall sentiment. If the compound score for a tweet is less than or equal to -0.05 then it is a negative sentiment and if it is greater than 0.05, then it shows a positive sentiment otherwise it is neutral. An example is shown below.

Table 2. Polarity scores of sample tweets

Tweet	Polarity Score
rt years of being called slow and here we are with Cardano and Ethereum at the same general level of core development	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Ethereum is going down. I am selling it	{'neg': 0.23, 'neu': 0.77, 'pos': 0.0, 'compound': -0.12}

The compound sentiment was the main focus of this paper. Sample tweets that were retrieved for a particular day was cleaned and passed through VADER's sentiment intensity analyzer in the NLTK library. The average compound sentiment for a day was then taken by dividing the total compound sentiment of tweets that day by the number of tweets collected for that day.

$$Average\ Compound\ Sentiment = \frac{Total\ compound\ sentiment}{Total\ number\ of\ tweets\ collected\ for\ the\ day} \quad (2)$$

4.3 Visual Exploratory Data Analysis

Distribution.

The plots of the features, which show the possible values for a variable and the frequency with which they occur, revealed that the distribution of sentiments followed a normal distribution, the distribution of tweet counts followed a normal distribution as well, and the distribution of trends followed a positively skewed distribution.

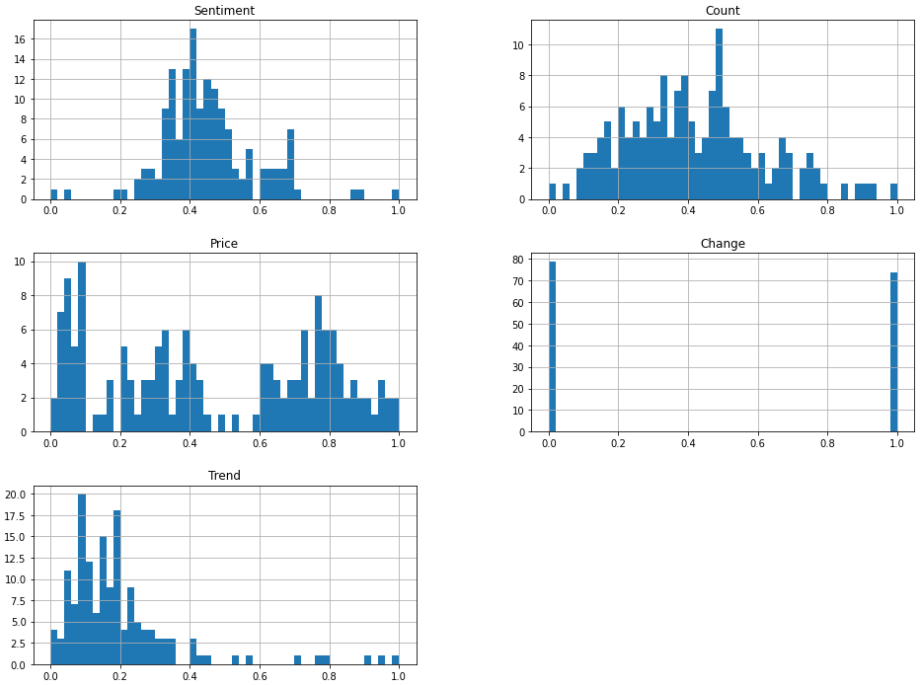

Fig. 2. Charts showing various distribution of the variables considered in this research.

Sentiment and Cryptocurrency prices.

The correlation was computed, and the result was 0.412, which is approximately 41%. This was done so that we could determine the extent to which the magnitude of emotion for a day impacts the price of Ethereum. According to what can be inferred from the figure below, the general attitude was positive on average, regardless of how much the

price of Ethereum moved.

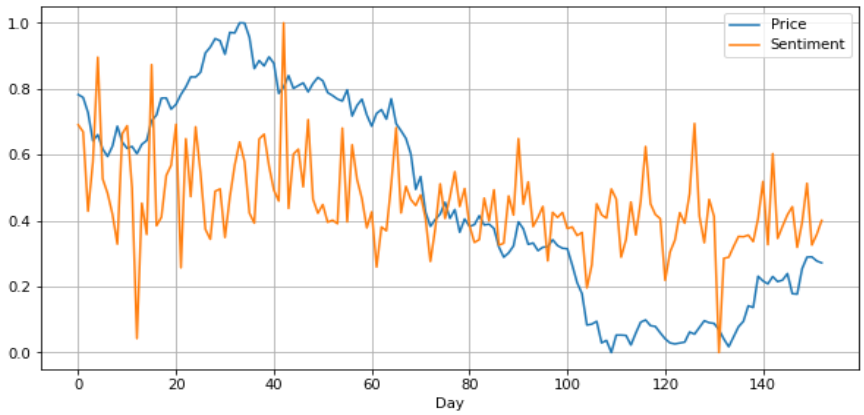

Fig. 3. A plot of sentiment compared to Ethereum prices

Tweet Count and Cryptocurrency prices.
Similarly, the correlation between tweet count and the cryptocurrency prices was also computed as -0.375506. This means that that an increase in price lead to a decrease in the number of tweets on Ethereum for that particular day.

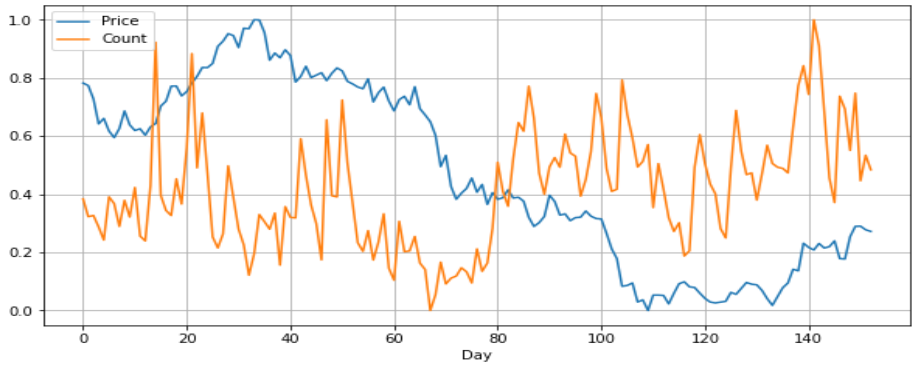

Fig. 4. Plot of tweet counts and Ethereum prices over 153 days

Trend and Cryptocurrency prices.

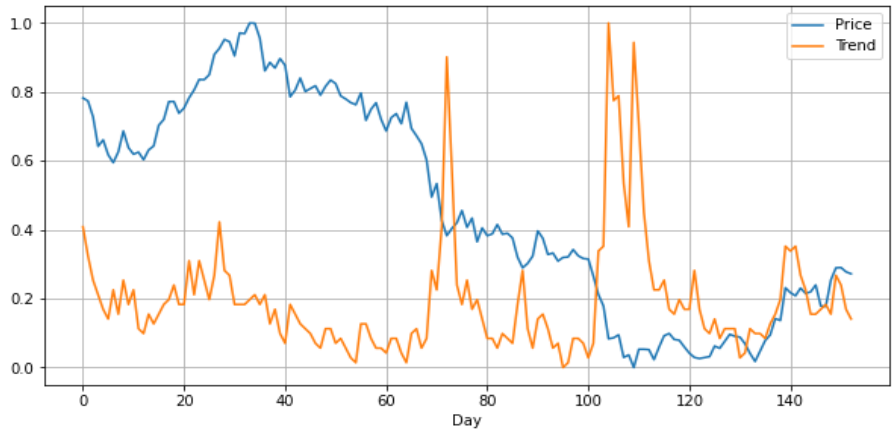

Fig. 5. Plot of Interest in Ethereum and Prices against 153 days.

5 Method

Two regression models were used to train a dataset containing four features: a sentiment, tweet count, trend and the label, market change which calculated using the open price and the close price. The main dataset was split into two, the testing and training set. The test set was 40% of the overall amount of data. Before training the models, I had the following sample of data in data.csv

Table 3. A sample of raw data from the main final dataset.

Sentiment	Count	Trend	Change
0.243190	91952	58	1
0.235499	86876	52	0
0.148117	87218	47	0
0.202503	83943	44	0
0.317556	80204	41	1

Table 4. Sample of dataset after normalization (Used for Training the models)

Sentiment	Count	Trend	Change
0.690604	0.383887	0.408451	1.0
0.669389	0.322770	0.323944	0.0
0.428362	0.326888	0.253521	0.0
0.578376	0.287455	0.211268	0.0
0.895729	0.242436	0.169014	1.0

5.1 K Nearest Neighbor

The K-Nearest Neighbors is supervised learning algorithms popularly used in machine learning for classifications. It makes use of proximity to create classifications or predictions about the grouping of an individual data point. It is also known as KNN or k-NN. Although it may be used to problems involving either regression or classification, it is more frequently used as a classification approach since it is based on the idea that points with similar properties can be found in close proximity. In K-Nearest Neighbors, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. The factor K indicates the number of neighbors to considered when classifying an input point.

In point of fact, the project utilized a KNN implementation called KNeighborClassifer class from scikit learn library in python. It's hyperparameter, n_neighbors was set to 8 after performing a hyperparameter tuning on the model as it provided the best results.

```
KNeighborsClassifier(n_neighbors=8)
```

The model was then fit to the training set; X_train and y_train and evaluated by testing on both the training and the test set. The final evaluation was done using cross-validation and taking the mean of the accuracy scores. The results are discussed in section 5.4.

5.2 Gaussian Naïve Bayes

Naïve Bayes classifiers are simply probabilistic models usually applied to classification tasks in machine learning.

The model was then fit to the training set. The final evaluation was done using cross-validation and taking the mean of the accuracy scores. The results are discussed in section 5.4.

5.3 Random Forest Classifier

Random Forest is an algorithm that consists of multiple decision trees and usually applied to classification tasks. It is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting a number of decision tree classifiers on various sub of the dataset. The number of trees in this project is fixed at 100.

```
RandomForestClassifier(n_estimators=100)
```

The model was then fit to the training set; The final evaluation was done using cross-validation and taking the mean of the accuracy scores. The results are discussed in section 5.4.

5.4 Results

After training all the models on the training set, the prediction was taken for both the training and test set. But more importantly, I considered evaluated and recorded scores from cross validation on the training set. For each model, the mean of cross-validation scores for both normalized and unnormalized data were recorded.
For the K Nearest Neighbor Classifier, I achieved an accuracy score of 0.42 on raw data and a model with accuracy as high as 0.57 on normalized data. This goes to show that scaling dataset influences model performance. Additionally, with Random Forest Classifier, unnormalized data also produced a model with an accuracy score of 0.51 and 0.51 when the data was normalized. Finally, for Naïve Bayes, the model did not really perform well. It produced a model with an accuracy of 0.42 and 0.38 for raw and normalized data respectively.
The best model in this project was able to accurately predict 57% percent of the time whether there is going to be an increase in price (1) or a decrease (0) given the sentiment, tweet counts and trend for a particular day.

6 Conclusion

I was able to demonstrate that twitter sentiment and volume is not a good predictor of Ethereum prices but a good indicator of market movements using both K Nearest Neighbor and Random Forest Classifier. When the dataset was normalized, I was able to attain an accuracy of approximately 57%, which was originally 42% with the raw data. This article also provides evidence that there is a relationship between the total number of tweets and market pricing. It was demonstrated that there was a slightly negative correlation between the price of Ethereum and the quantity of tweets, and vice versa.

There is still a lot of work to be done in this area. The use of sentiment analysis for gathering social signals might be improved by increasing the quality of the information and the number of sources from which it is obtained. Quality might be improved by removing duplicate information and screening out tweets by bots and ads. Another area of opportunity is the use of more specialized models with different types of attention mechanisms, such as long short-term memory networks (LSTM. Recent research has shown that the predictability of LSTMs is significantly higher when compared to generalized regression neural architecture (Lahmiri & Bekiros, 2019). These networks may be able to detect the market's intrinsic "moods" and adjust accordingly.

In addition, I recommend that distinct models be used for the data obtained from Twitter and the market in order to get higher accuracy and precision ratings for the models. It would be fascinating to finally demonstrate whether or not these prediction models can be utilized for the development of trading strategies.

References

Critien, J. V., Gatt, A., & Ellul, J. (n.d.). *Bitcoin price change and trend prediction through twitter sentiment and data volume*. https://doi.org/10.1186/s40854-022-00352-7

Hootsuite. (2022). *33 Twitter Statistics That Matter to Marketers in 2022*. Hootsuite's 2022 Digital Trends Report. https://blog.hootsuite.com/twitter-statistics/

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons and Fractals*, *118*, 35–40. https://doi.org/10.1016/j.chaos.2018.11.014

Mittal, A., & Goel, A. (n.d.). *Stock Prediction Using Twitter Sentiment Analysis*.

Rao, T., & Srivastava, S. (2012). Analyzing Stock Market Movements Using Twitter Sentiment Analysis. *Undefined*. https://doi.org/10.1109/ASONAM.2012.30

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy 2019, Vol. 21, Page 589*, *21*(6), 589. https://doi.org/10.3390/E21060589