Bitcoin trend prediction through twitter sentiment*

Cheng, I-Chun
School of Computing
National University of Singapore
Singapore
e0957058@u.nus.edu

Dong Zheng
School of Computing
National University of Singapore
Singapore
e0450318@u.nus.edu

Evelyn Chen
School of Computing
National University of Singapore
Singapore
e0313687@u.nus.edu

Qizhuang Li
School of Computing
National University of Singapore
Singapore
e0957022@nus.edu.sg

Xie Jialong
School of Computing
National University of Singapore
Singapore
e0332691@u.nus.edu

ABSTRACT

Twitter sentiment analysis has been shown to be useful for predicting Bitcoin trends. In this paper, we seek to build on the state-of-art to predict the trend of Bitcoin in terms of increase and decrease in price, with the assumption that sentiment expressed is a reliable indicator of price change within a window of time. We collected a dataset of Bitcoin-related tweets using the Twitter API and performed sentiment extraction on the tweet content. Two neural network models are explored and evaluated, one based on a variety of recurrent neural networks (RNN) and one based on the convolutional network (CNN). An additional binary-classification model is implemented and compared with the neural network models. Our results show that the RNN based model yields a higher accuracy in predictions when used alongside the other two models. The main outcome of this project demonstrates the relationship between sentiment and Bitcoin price trend, and the trend can be predicted with machine learning models with a relative accuracy of (55%).

KEYWORDS

Bitcoin, prediction models, sentiment analysis, PageRank, MapReduce

1 Introduction

The launch of Bitcoin in around 2008 has attracted great attention from around the world. Because of its features- Immutability, transparency and can be transacted through decentralized networks, more people use it to trade, invest, and some may use it to store personal property. As it becomes one of the mainstream decentralized currencies, its change in price influences a lot of investors and speculators, as well as other Bitcoin-denominated products. Therefore, many of them have attempted to predict the changes in Bitcoin price using various methods.

The fanaticism of Bitcoin has aroused tens of thousands of discussions in all kinds of social media platforms. Countless forums and posts about Bitcoin are produced every day, and some may include the prediction and interpretation of the future tendency, while others just express their emotions on the platform. Believing people would directly share their feelings on social media websites, in this research, we capture Twitter's comments related to Bitcoin, categorize it based on the tweets' emotions, and analyze this data to see its correlation with Bitcoin's price changes.

2 Data Preparation

2.1 Bitcoin Price Data

To study the Bitcoin price movement, we first extracted the Bitcoin price from yahoo finance, which requires Python yfinance - an open-source tool to download financial data. We extracted "BTC-USD" data with columns of "Open", "High", "Low", "Close" and "Volume", and further calculated the percentage change of the "Close" price. With this percentage change, we can label the data of each period with the value of either "UP" or "DOWN" to indicate the binary movement of the Bitcoin price.

2.2 Bitcoin Tweet Data

We tried to scrape bitcoin tweets from Twitter using twint - a popular tweets scraping tool. However, due to some internal bugs within the tool as also posted and reported by others, we were not able to extract all tweets with the specified dates.
Alternatively, we used sntwitter from snscraper to scrape tweets with desired dates. To remove random posts about Bitcoin and

minimize the noises, we chose to extract Bitcoin tweets with 10 or more likes to ensure that the tweet sentiments we used are somewhat agreed by some other people. The key data variables we need for sentiment analysis are username, tweet content, datetime and the number of likes. In total, we scraped around 500,000 tweets for later sentiment analysis.

2.3 Sentiment Analysis

Before sentiment analysis, tweet contents were cleaned by removing duplicates, URLs, emojis, hashtags, stopwords, mentions and punctuations.

Cleaned contents would be sent to sentiment analysis after lemmatization. Using SentimentIntensityAnalyzer, scores of positive, neural, negative and compound sentiments are created for each tweet. Positive/neutral/negative scores range from 0 to 1 and compound score ranges -1 (to indicate the most negative sentiment) to 1 (to indicate the most positive sentiment).

To visualize the existence of the effect of tweets sentiment on Bitcoin price, we created a heatmap as shown in Figure 1 below. The positive and negative sentiments show relatively significant correlation with the close price of Bitcoin.

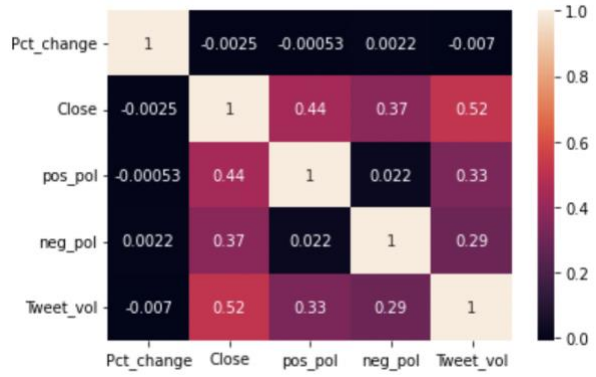

Figure 1: Correlation Matrix

2.4 Data Combination

Tweet sentiment data and bitcoin price are aligned and combined based on time period. In our project, we used data of 3 different time frequencies - every 30 minutes, hourly and daily.

3 Twitter PageRank

In this section, we will discuss how we find the top influencers in the Twitter Bitcoin community, and how we cluster into different community.

3.1 Generate Directed Graph from Data

Twitter is a place where people can post their opinions about interesting things. In addition to this blog-like functionality, twitter also allows users to interact with each other by using mention and retweet functions. This makes the twitter community structured like a graph, each user can be considered as a node and the mentions can be treated as hyperlink or edges between nodes. Therefore, we decided to use scraped tweet data mentioned in the previous section to construct the network of the BTC community on twitter.

We used regular expressions in the user defined function of PySpark to exact mention information in each tweet. If there is a mentioned username in the tweet, a directed edge will be formed from the host twitter account to this mentioned account, and we finally get this table containing edge information in the all tweet data.

	source	target	Timestamp
0	@VeronicaLake21	@BTC_for_Freedom	2022-10-23 02:35:06+00:00
1	@VeronicaLake21	@ZERP589	2022-10-23 02:35:06+00:00
2	@VeronicaLake21	@XRPrincesses	2022-10-23 02:35:06+00:00
3	@BTC_for_Freedom	@Croesus_BTC	2022-10-12 16:57:31+00:00
4	@TIPMayerMultple	@TIPMayerMultple	2022-10-11 15:01:00+00:00
...	...	...	...
208069	@Cryptotigers8	@CryptoWendyO	2022-10-11 13:15:48+00:00
208070	@tzongocu	@binance	2022-10-11 13:15:35+00:00
208071	@FonsiKristen	@CharityToken4	2022-10-11 13:15:27+00:00
208072	@xrp_mxracing	@BCBacker	2022-10-11 13:15:16+00:00
208073	@MillionIwanta	@cryptojack	2022-10-11 13:15:06+00:00

208074 rows × 3 columns

Figure 2: Twitter User Mention Table

3.2 Find Influencers by using PageRank

The total number of nodes in this directed graph is 104,573, and the total number of edges is 208,074. Then we use networkx's pagerank algorithm with a damping parameter of 0.85 to calculate the pagerank score of every twitter account in this graph. The sorted result of top influencer in the twitter bitcoin community is shown in the table below:

Table I: Top influencers' PageRank

Ranking	Account Name	PageRank Score
0	@BTC_Archive	0.014991869
1	@elonmusk	0.007077413
2	@deezy_BTC	0.006038315
3	@Croesus_BTC	0.005124831
4	@BTC_for_Freedom	0.00424186
5	@rovercrc	0.00398958
6	@btc_charlie	0.003900807
7	@saylor	0.003567192
8	@MicroStrategy	0.003423356
9	@ptc_sitapur	0.002612067
10	@binance	0.002528573
11	@Trainwreckstv	0.002324842
12	@TheMoonCarl	0.002088679
13	@boltcoiner	0.002087244
14	@_bitcoiner	0.002044377
15	@AltCryptoGems	0.00191513

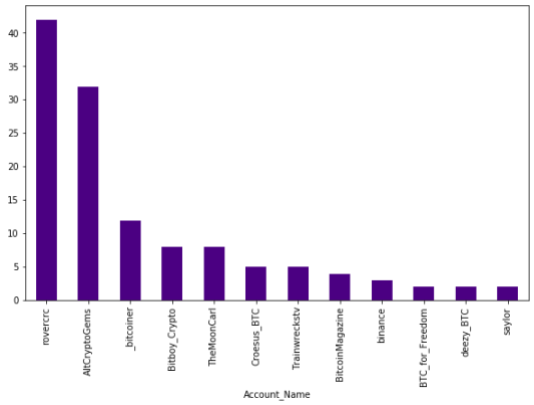

Figure 3: Number of tweets by Top Influencers

We also count the number of tweets posted by these influencers' accounts, and find that top influencers do not necessarily have most tweets posted on the internet. For example, @elonmusk does not have a single tweet mentioning BTC in our scraped database, but he is still the second most influential person in the tweet BTC community, because many other users mentioned him in their tweets. This result is aligned with the flow model, the account's own importance is the sum of the in-links votes.

3.3 Community Detection

We would also like to identify if the Bitcon community tends to cluster in groups or scattered around. This will imply whether they share the same or different opinions about BTC, and the result will be useful for price prediction in the next step. In order to detect communities in this twitter network, we first need to identify individual's betweenness centrality. Betweenness centrality measures the number of shortest paths passing through a node. In this case, an account with high betweenness centrality means it has a high probability not belonging to any community and we can use it to split the graph. Meanwhile, good communities means there should be enough intra-community edges. By using these two criterias, we were able to identify communities in out dataset as below:

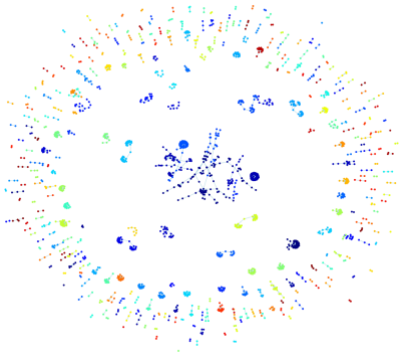

Figure 4: BTC Community Detection

3.4 Future Improvement

When calculating pagerank score and plotting the community graph, I found that there is a potential spamming structure between @boltcoiner and @_bitcoiner. @_bitcoiner is the spammer and @boltcoiner creates lots of farmer tweets mentioning @boltcoiner so that their pagerank score increases drastically. By noticing this, we remove their tweets from BTC price analysis in the next step.

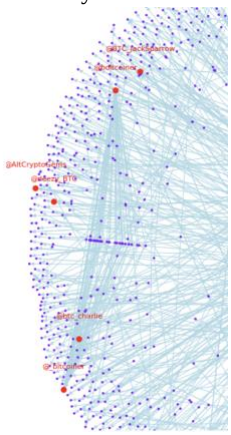

Figure 5: Potential Spamming Structure

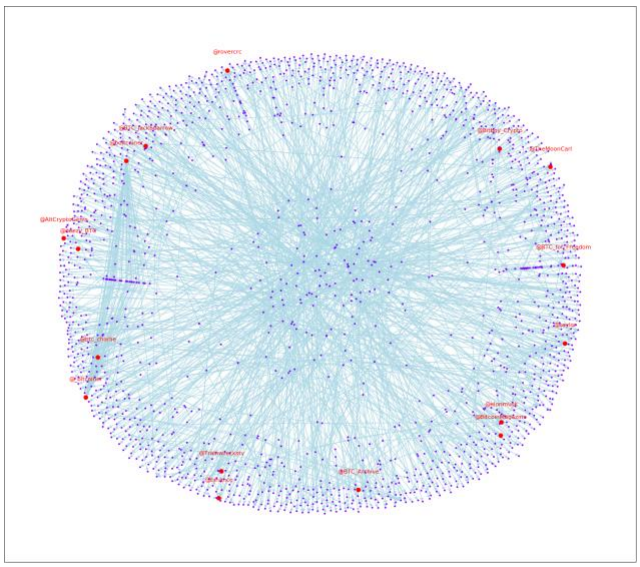

Figure 6: Overall BTC Twitter Network

We also propose a way to improve this problem by using TrustRank. Basically we will further collect information about whether a twitter account is verified or not, and use them as trusted seed to propagate trust among twitter BTC community. If an account's trust score is below some threshold, we will not include theri tweets in the further analysis.

4 Data Analysis

4.1 Linear Regression

We performed a linear regression with sentiment score as the independent variable and percentage change as the dependent variable. The Hypothesis H0 was: There is no linear relationship between sentiment score and percentage change.

The scatter plot along with the trend line are shown in Figure 7, with sentiment score on the x axis and percentage change on the y axis.

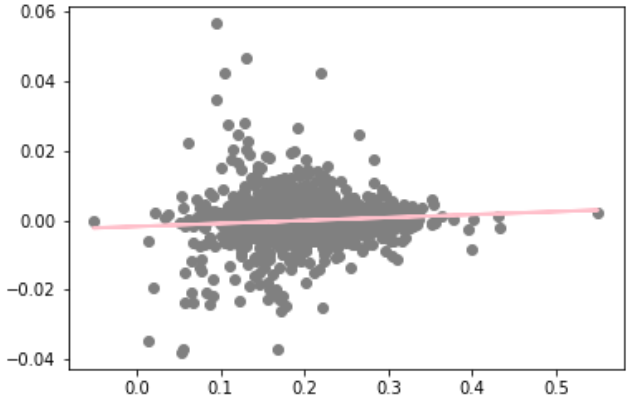

Figure 7: Linear Regression Model Plot

As the p-value is 0, we can reject the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). With the confidence that the sentiment score is positively correlated with the percentage change, we continue to perform the following tests.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.005
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	31.10
Date:	Sun, 30 Oct 2022	Prob (F-statistic):	2.56e-08
Time:	14:16:07	Log-Likelihood:	20304.
No. Observations:	5637	AIC:	-4.060e+04
Df Residuals:	5635	BIC:	-4.059e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0017	0.000	-5.750	0.000	-0.002	-0.001
x1	0.0081	0.001	5.577	0.000	0.005	0.011

Omnibus:	1029.489	Durbin-Watson:	1.889
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26509.680
Skew:	0.074	Prob(JB):	0.00
Kurtosis:	13.623	Cond. No.	17.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 8: Linear Regression Results

4.2 Binary Logistic Regression

Since we aim to predict whether the bitcoin trend is going up or down in the next time interval, the dependent variable y is a binary output. It makes more sense to use a binary logistic regression to predict the outcome based on the sentiment score. The Binary Logistic Regression regresses to the following equation.

$$\log(\frac{p}{1-p}) = b_0 + b_1 X$$

In this equation, p is the probability that the trend goes up given the sentiment score, where the trend is the dependent variable and sentiment score is the independent variable. b_0 and b_1 are the parameters of the model, which are estimated using the maximum likelihood method. The left-hand side of the equation ranges between minus infinity to plus infinity.

The Hypothesis H_0 was: There is no relationship between sentiment score and percentage change.

From Figure 4 below, as the p-value is 0.001, we can reject the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random).

Logit Regression Results

Dep. Variable:	Pct_change	No. Observations:	5637
Model:	Logit	Df Residuals:	5635
Method:	MLE	Df Model:	1
Date:	Sun, 30 Oct 2022	Pseudo R-squ.:	0.001481
Time:	13:39:17	Log-Likelihood:	-3901.3
converged:	True	LL-Null:	-3907.1
Covariance Type:	nonrobust	LLR p-value:	0.0006706

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.3089	0.090	-3.420	0.001	-0.486	-0.132
sentiment_score	1.5022	0.442	3.395	0.001	0.635	2.369

Figure 9: Binary Logistic Regression Results

The Binary Logistic Regression gave the following confusion matrix, resulting in an accuracy of 0.510732659215895.

Table II: Confusion Matrix of Binary Logistic Regression

	Actual Up	Actual Down
Predicted Up	1661	1180
Predicted Down	1578	1218

The accuracies of the prediction models are calculated using the following formula:

$$\text{Accuracy} = \frac{True\ Ups + True\ Downs}{Total\ Number\ of\ Predictions}$$

5 Methodology

The prediction of cryptocurrency price trend can be formulated as a binary-classification problem, where the task is to predict up/down of the price at a specific timeframe given corresponding feature inputs. The input features are:

1. Close price
2. Vader polarity score - positive score
3. Vader polarity score - negative score
4. Tweet Volume (group by specific timestamp)

Label is pct_change, which is calculated as the percentage change between the current and the prior price given a specific timeframe.

5.1 BiLSTM Model

The Long-Short Term Memory (LSTM) network is a variety of the recurrent neural network that is well-suited to classifying, processing and making predictions based on times series data, since there can be lags of unknown duration between important events in a time series. BiLSTM is a special type of LSTM that runs the inputs in two ways, one from the past to future and one from future to past. Therefore, BiLSTM can have information on both past and future compared to traditional LSTM models. The model implemented is actually CuDNNLSTM -- a fast LSTM implementation backed by CuDNN that can only be run on GPU, with the TensorFlow backend. The architecture of this model is depicted in Fig 10.

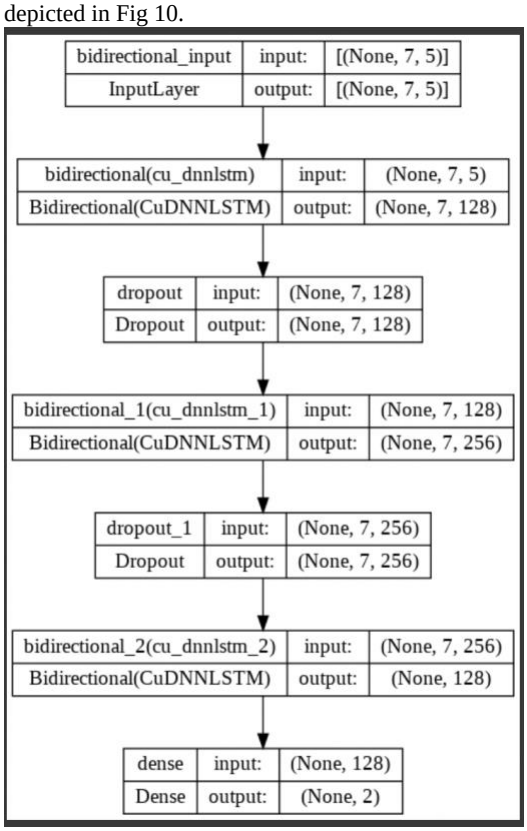

Figure 10: BiLSTM model architect

5.2 CNN Model

CNN's convolutions are popularly known to work on spatial or 2D data. However, there are also convolutions for 1D data. This allows CNN to be used in more general data types including texts and other time series data. Instead of extracting spatial information, 1D convolutions can be used to extract information along the time dimension. The architecture of this model is depicted in Fig 11.

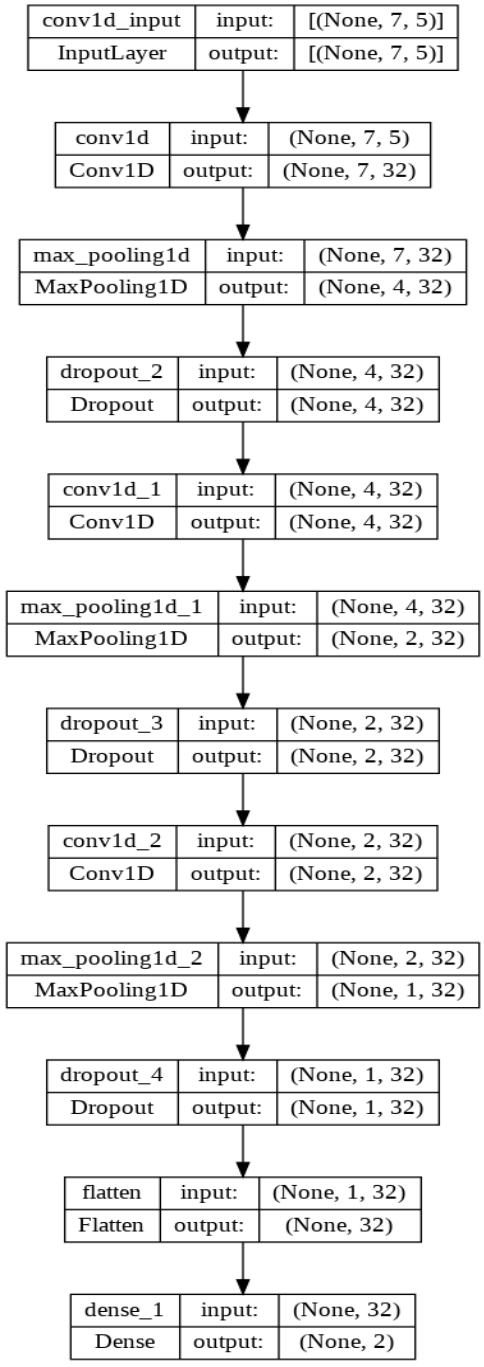

Figure 11: CNN model architect

5.3 RNN-based Models vs. CNN-based Models

1) Computational Complexity: CNNs are computationally cheaper than RNNs. CNN learns by batch while RNNs train sequentially. As such, RNNs can't use parallelization because they must wait for the previous computations.

2) Assumptions on Data: CNNs do not have the assumption that history is complete. Unlike RNNs, CNNs learn patterns within the time window. Therefore, CNNs would outperform RNNs in the case of missing data.

3) Data Scanning: RNNs only learn from data before the timestamp it needs to predict. However, CNNs (with shuffling) can observe data from a broader perspective.

6 Results

When the data were aggregated by hour, the accuracies of the three models are shown in Figure 12.

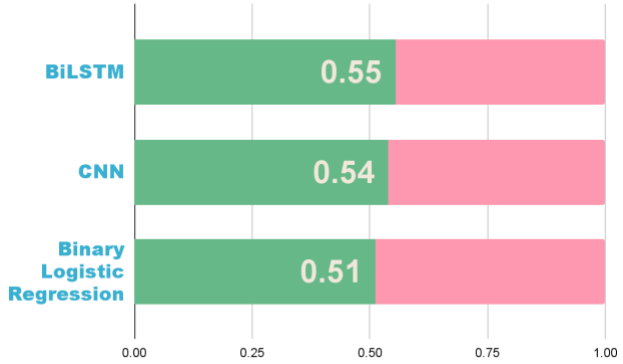

Figure 12: Accuracy using Hour Data

However, when the data were aggregated by day, the accuracies improved for all three models, as can be seen from Figure 13. The CNN model has the highest accuracy.

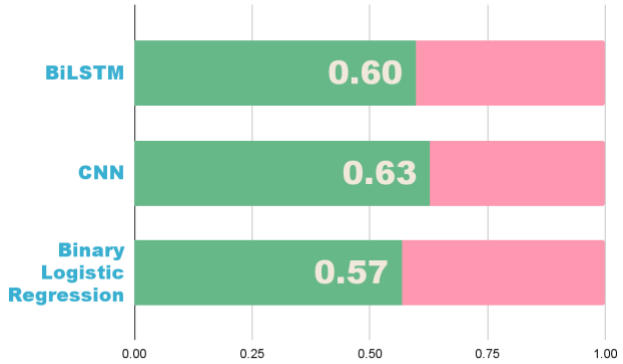

Figure 13: Accuracy using Day Data

7 Conclusion and Future Work

In this paper, we built on the state-of-art to predict the trend of cryptocurrency in terms of increase and decrease in price using sentiment expressed on twitter as an indicator of price change within a window of time. We first collected a dataset of crypto-related tweets using the Twitter API and performed sentiment extraction on the tweet content. We then proved our assumption that there is a correlation between the sentiments and the crypto price trends. Two neural network models are used and compared, giving a relative accuracy of 55%.

There are several interesting options for future work. We could use the result from PageRank, to weigh the sentiment scores, then pass the adjusted sentiment scores as an input to the models to predict the trend. We could also combine the results with trading strategies using technical indicators such as MACD or RSI to enter the market. We will definitely have to perform backtesting to evaluate our trading strategy, but we also need to bear in mind that the past is not representative of the future, and successful past performance may not guarantee that we will earn money.

REFERENCES

[1] Critien, J.V., Gatt, A. & Ellul, J. Bitcoin price change and trend prediction through twitter sentiment and data volume. Finance Inov 8, 45 (2022). https://doi.org/10.1186/s40854-022-00352-7

[2] Manojit, N. Social Network Analysis with NetworkX. (2015). From https://www.dominodatalab.com/blog/social-network-analysis-with-networkx

[3] James A. Extracting Interactions Networks from Twitter using TWINT and Python (2021) from https://dataground.io/2021/06/25/extracting-interactions-networks-from-twitter-using-twint-and-python/

[4] Abraham J, Higdon D, Nelson J, Ibarra J (2018) Cryptocurrency price prediction using tweet volumes and sentiment analysis

[5] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on language in social media (LSM 2011), pp 30–38

[6] Baker M, Wurgler J (2007) Investor sentiment in the stock market. J Econ Perspect 21(2):129–152