SINGAPORE MANAGEMENT UNIVERSITY

ISSS606 – Social Analytics

Group 8

Project Report

Names: Jordan Ong Zhi Rong
Kenneth Low Yan Wei
Peace Tay Jiunn Ching
Tan Yu Yan, Rachel
Widya Tantiya Yutika
Yap Pin Yaw

Contents

List of Figures

List of Tables

1. Background

Prices of traditional stocks are affected by factors such as a company's fundamentals, macroeconomic and industrial factors, and prospects. Conversely, cryptocurrency's prices are driven primarily by supply and demand. Therefore, cryptocurrency's price changes are more volatile and susceptible to public's opinion.

This project aims to assess the correlation between social sentiments, and importance of Twitter users' profiles on its impact on cryptocurrency's price change. More specifically, our analysis will be focused on user influence, such as the number of followers and tweet sentiment compound scores as the main components of measuring popularity. To extend our scope for a more comprehensive overview, discussions of cryptocurrency on Reddit are also used to complement our analysis of the general population on a different social media platform.

2. Dataset

Bitcoin, the first decentralised cryptocurrency built upon the blockchain technology, was introduced in 2009 by an anonymous developer who goes by the name of Satoshi Nakamoto. With the success of Bitcoin, interest in blockchain technology grew and many cryptocurrencies were created (The Investopedia Team, 2020).The fundamental reason for choosing Bitcoin for this project is because it is the most well-known cryptocurrency and its dominance of the cryptocurrency market capitalization (>50%), which affords Bitcoin a lot of influence over the price behaviour of other coins (Chipolina, 2020). Therefore, we will use Bitcoin as a proxy study subject for the cryptocurrency market.

Variable	Description	Source
Bitcoin	• Historical pricing data	www.investing.com
Influencer & Institution Tweets	• Tweets of selected individuals, organizations, and the general population	Twitter API Snscrape
Bitcoin in Reddit Community	• Posts in Reddit's Cryptocurrency community	Pushshift Reddit API

In this project, we chose a list of users (both influencers and institutions) we hypothesize as the most influential regarding cryptocurrency interest on the Twitter social platform. Tweets and discussions on Reddit by the general population were also considered in this study.

Influencers	
Elon Musk	• He has ~ 98.9 million followers on Twitter where he influences the crypto market, leading to the coining of the term "Musk Effect".
Changpeng Zhao	• Founder of Binance, has ~ 6.4 million followers on Twitter, he has grown to become the largest cryptocurrency exchange for retail investors in the world.
Nayib Bukele	• The President of El Salvador has ~ 4.0 million followers on Twitter. During his presidency, El Salvador became the first country to accept Bitcoin as legal tender.
Vitalik Buterin	• The co-founder of Ethereum, has ~ 4.0 million followers on Twitter, where he shares insightful commentaries about crypto and blockchain.

Institutions	
Bloomberg	• Bloomberg is the global leader in business and financial data, news, and insight. • Bloomberg is widely followed by institutional, retail investors, and market participants and has a total of 8.3M followers in twitter
Goldman Sachs	• It is the first big U.S. bank to offer its wealthy clients access to Bitcoin funds. • 940K followers in Twitter

General Population	
Twitter Tweets	• All tweets about bitcoin from any twitter handler

Reddit Posts	• All the posts in the bitcoin subreddit which has 4,300,000 members with average of 3,600 active online.

2.1 Data Processing

2.1.1 Data Collection

For data collection, our team scrapped an estimated of 10 million tweets from Twitter (over 3 years) using *snscrape* library and estimated 50,000 Reddit posts (over 3 years) using *psaw* library, and *pushshiftAPI*. To prevent any data bias during the collection of data, no additional filters (only BTC, Bitcoin) were used during scrapping of text for general population and influencers/institution.

2.1.2 Pre-Processing

All tweets and posts were pre-processed to ensure relevancy of text and achieve a better result from the analysis. Our team leveraged on existing libraries such as *NLTK, langdetect* and *Regex* to pre-process the tweets. The pre-processing is conducted in 3 different steps: **Context Cleaning**, **Valid Word** and **Language Detect**. For Context Cleaning, the strategy was to convert twitter or reddit context words such as RT (retweet), URL, @ (mention) and emojis using Regex to enhance the effectiveness of the sentiment analysis. For Valid Word, define valid words were defined as words that consist of letters from A-Z and numbers from 0-9, it was also defined that words with more than 2 consecutive repeated letters are invalid and are reduced to 2 (e.g. “funnny” will be converted to “funny”). Lastly, for Language Detect, tweets that are non-English were removed.

#	Steps	Description	Tools/Library used
1	Scraping Twitter and Reddit	Scaping documents	TwitterSearchScraper, snscrape, Pushshift Reddit API
2	Filtering documents	Filter tweets related to cryptocurrency	Regex
3	Convert documents to lower case	Document cleaning	Regex
4	Replace URL, @ Handles and # Hashtags		Regex
3	Remove Retweets and Tweets with <3 words		Regex
4	Remove Spaces, Emojis ‘		Regex
5	Remove stopwords and stemming		PorterStemmer
6	Remove punctuations, - and & and repeated letters	Valid word checking	Regex
7	Check letters and numbers		Regex
8	Check for English tweets	Filter English tweets	LangDetect

3. Tools and Environment

A set of Python libraries were utilised, in collecting, processing, analysing, and displaying data. The major modules that were used are:

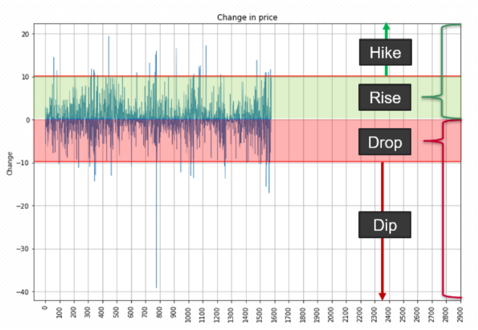
- **Snscrape**. It is a scraper from Python library that can be used to scrape data through Twitter’s API without any restrictions or request limits. However, not all data can be retrieved using snscrape such as follower’s ids.
- **Twitter API**. It is a set of programmatic endpoints allow users to scrape various data from twitter.
- **Pandas.** It is a Python library that utilises the fast and efficient structure of DataFrames with automatic initialisation indexes and offers data alignment.
- **Pushshift Reddit API.** It was created by the /r/datasets mod team to help provide enhances functionality and search capacilities for searching Reddit comments and submissions. Pushshift Reddit API is used because it can be filtered by timeline.
- **NLTK.** It is used for analyzing, preprocessing and understanding written text i.e. sentiment analysis.
- **Networkx.** A Python library for studying graphs and networks.
- **Regex.** It is used to check if a string contains a specified search pattern.

4. Approach

The following sections will describe the approach taken to conduct the below analyses:

- Social Influence
- Topic Modelling
- Sentiment Analysis

Two methods were devised to explore correlation between social sentiments, and topic modelling with respect to price changes. Firstly, the relationship between daily price changes and sentiment scores were explored. Secondly, to investigate the reasons for significant price changes, the following definitions were used to define these extraordinary days for further analysis.



- **Price Hike Days:**
  Days with > +2.5 standard deviation price change from mean (>10%)
- **Price Rise Days:**
  Days with any positive price change (>0%)
- **Price Dip Days:**
  Days with > -2.5 standard deviation price change from mean (< -9.8%)
- **Price Drop Days:**
  Days with any negative price change (<0%)

Figure 1: Price Change (%) versus Days (SN)

4.1 Social Influence

From the each of the selected four (4) influencers and two (2) institutions identified, 5,000 followers were extracted, resulting in a total of 30,000 follower nodes. After extracting the number of followers that each of the 30,000 nodes have, the list was sorted in descending order, to identify the top 30 users (based on number of followers). Another 5,000 followers were extracted from the top 30 users identified, resulting in a total of 180,000 follower nodes.

From the nodes obtained, the below proxies will be used to reflect the six (6) influencers' & institutions' influence:

- **Network Visualisation.** The nodes were simplified by removing any nodes with less than 2 degrees, to ensure significant contribution of nodes in the network graph.
- **PageRank Centrality.** PageRank scores of each of the influencers and institutions were also computed.

4.2 Sentiment Analysis

The project attempted to analyse whether sentiment scores of influencers and the general population on social networks (i.e., Twitter and Reddit) are correlated to cryptocurrency prices, and whether it could be regarded as a predictive premise for the increase and/or decrease of cryptocurrency prices. For the analysis, we will be adopting the use of confusion matrix for the individual influencers, and the strength of correlation between the general population of Twitter and Reddit with the price changes of cryptocurrency.

NLTK VADER sentiment analyser was employed to retrieve the compound scores for tweet/post relating to cryptocurrency. The daily average score for the tweets and posts were used in analysis. For twitter, the respective compounds scores of each of the influencers' tweets as well as the general population tweets were retrieved. For reddit, only the scores of the general population's posts were retrieved.

As part of the analysis, individual influencers' daily tweet compound score was compared with the price change of the day for bitcoin. The findings were translated to an illustration via a confusion matrix. Regarding price change, two aspects were considered, one aspect where only when there were significant price changes (i.e., hike/dip), and the other aspect considers whenever there was a price change.

4.3 Topic Modelling

There exists a possibility that price changes in cryptocurrencies are affected by factors other than public sentiments of the coin. To further explore what topics are being discussed on days with significant price changes (i.e. Price Hike and Price Dip), Topic Modelling was conducted.

After pre-processing was done to obtain tokenized tweets, *Latent Dirichlet Allocation (LDA)* was applied and *Gensim* was used to obtain the topics. After tuning, 10 topics were found to produce the most coherent set of topics, and the top 5 words in each topic were set for the parameters of this model to be printed using *WordCloud* and *matplotlib.pyplot*.

5. Results

The following sections will describe the results obtained from the below analyses:

- Social Influence
- Topic Modelling
- Sentiment Analysis

5.1 Social Influence

5.1.1 Network Visualisation

To visualise the number of followers, the simplified list of nodes was used, resulting in the diagram below.

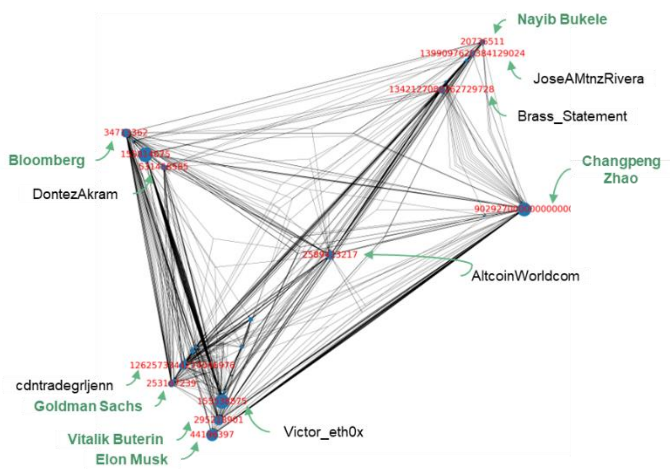


Figure 2: *Simplified Network Visualisation using Networkx*

The simplified in-degrees of the six (6) influencers' & institutions' were calculated, giving:

Influencers	
Elon Musk	470
ChangPeng Zhao	616
Vitalik Buterin	302
Nayib Bukele	100

Institutions	
Bloomberg	252
Goldman Sachs	144

Table 1: Simplified In-Degree Counts of Influencers & Institutes

Other nodes with significant resultant in-degree counts (above 100 in-degrees) were also identified:

User_ID	**Specialty / Focus**	**Simplified In-Degrees**
Brass_Statement	Defi payment network	182
JoseAMtnzRivera	Co-founder of XoloToken	102
Dontez Akram	Entrepreneur	112
Victor_eth0x	Cryptocurrency promoter	671
AltcoinWorldcom	Altcoins promoter	251
cdntradegrljenn	Equity Investor	162

Table 2: Simplified In-Degree Counts of Other Significant Nodes

From the results obtained, the below observation was made:

- **People of similar interests, follow similar accounts.**
  Upon looking into the focus of the nodes with significant simplified in-degree count, it was discovered that they share the same focus: Investing and Cryptocurrency. This serves to verify that the sampled followers are indeed invested in the cryptocurrency scene or are at least interested in the topic.

5.1.2 PageRank Centrality

PageRank scores of each of the six (6) influencers & institutions were calculated and visualised in the chart below.

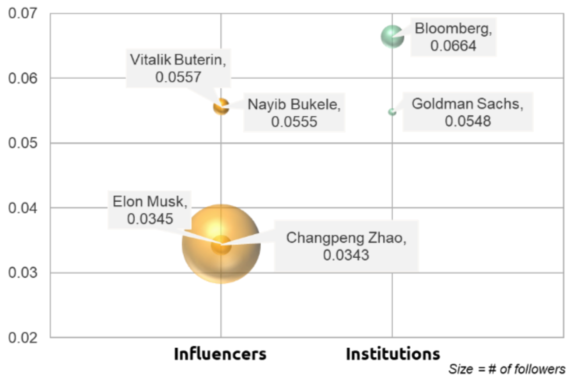


Figure 3: Visualisation of PageRank Scores of Influencers & Institutions

Number of Followers (millions):

Influencers	
Elon Musk	98
ChangPeng Zhao	6.4
Vitalik Buterin	4
Nayib Bukele	4

Institutions	
Bloomberg	8.3
Goldman Sachs	0.94

Table 3: Follower Counts of Influencers & Institutions

From the results obtained, the following observations were made:

- **Elon Musk and ChangPeng Zhao share a similar PageRank score.**
  From limited number of follower nodes sampled, a higher proportion of high-importance followers may have been sampled for ChangPeng Zhao, while a higher proportion of low-importance followers may have been sampled for Elon Musk, resulting in a similar PageRank score. Higher importance followers are defined as a person with high number of followers, but a low number of following.

- **Bloomberg has the highest PageRank score amongst the six (6) Influencers and Institutions.**
  Out of the group of six (6) chosen for the study, Bloomberg has the highest PageRank score, despite having significantly fewer followers than Elon Musk. This phenomenon can be alluded to the fact that there is no causal relationship between follower count and PageRank scores. It is possible that more retail investors and speculators tend to follow influencer accounts such as Elon Musk, resulting in a lower PageRank score. As compared to finance professionals who may tend to follow more important accounts such as new channels and renowned financial institutions such as Bloomberg, resulting in a higher PageRank score.

5.2 Sentiment Analysis

For sentiment analysis of influencers and institutions, utilizing the same definition for price hike/dip and price rise/drop, we will use the following confusion matrix for our analysis. Results between days with significant prices changes vs days with no significant price changes were compared.

	Actual Label = 0 (Price Dip/Drop)	Actual Label = 1 (Price Hike/Rise)
Predicted Class = 0 (-ve Compound Score)	True Neg (TN)	False Neg (FN)
Predicted Class = 1 (+ve Compound Score)	False Pos (FP)	True Pos (TP)

Figure 4: Confusion Matrix for Sentiment Analysis

With the confusion matrix, we derived each of the influencer's respective accuracy, precision, recall, and F1 scores. The results are illustrated in table form with visual representation in the figures below. The recall and F1 scores for Nayib Bukele and Goldman Sachs produced scores of 1 because the sample size of tweets happening during price hike/dip days were too little.

Elon Musk

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	1	24	2	33
Predicted Class = 1 (+ve Compound Score)	2	56	10	65

Changpeng Zhao

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	2	62	2	72
Predicted Class = 1 (+ve Compound Score)	4	222	9	252

Bloomberg

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	70	2295	82	2492
Predicted Class = 1 (+ve Compound Score)	119	2854	104	3138

Nayib Bukele

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	0	9	0	8
Predicted Class = 1 (+ve Compound Score)	2	27	3	33

Vitalik Buterin

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	4	174	9	185
Predicted Class = 1 (+ve Compound Score)	6	440	20	548

Goldman Sachs

	Actual Label = 0		Actual Label = 1	
	Dip	Drop	Hike	Rise
Predicted Class = 0 (-ve Compound Score)	0	8	0	8
Predicted Class = 1 (+ve Compound Score)	1	39	1	48

Influencers	Elon Musk		Nayib Bukele		Vitalik Buterin		Changpeng Zhao	
	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)
Accuracy	0.73	0.5	0.6	0.545	0.615	0.536	0.647	0.576
Precision	0.833	0.537	0.6	0.55	0.769	0.555	0.692	0.531
Recall	0.833	0.663	1*	0.805	0.690	0.748	0.818	0.778
F1-Score	0.833	0.593	1*	0.653	0.727	0.637	0.750	0.631

Institutions	Goldman Sachs		Bloomberg	
	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)	Sig. Price Change (i.e. Hike/Dip)	Price Change (i.e. Rise/Drop)
Accuracy	0.5	0.544	0.464	0.504
Precision	0.5	0.552	0.466	0.524
Recall	1*	0.857	0.559	0.557
F1-Score	1*	0.671	0.508	0.540

Figure 5: Confusion Matrix Scores for Influencers and Institutions

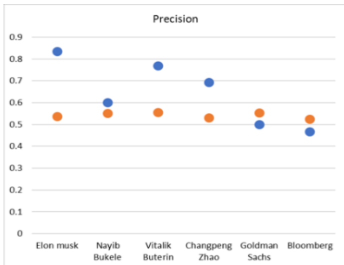


Figure 6: Precision Scores of Influencers and Institutions

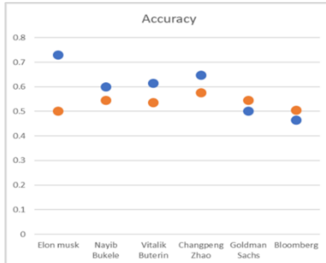


Figure 7: Accuracy Scores of Influencers and Institutions

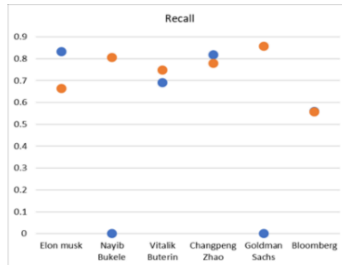


Figure 8: Recall Scores of Influencers and Institutions

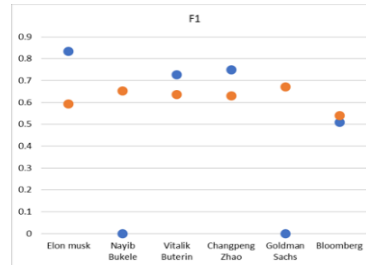


Figure 9: F1 Scores of Influencers and Institutions

** Blue dots represent price hike/dip, and orange dots represent price rise/drop*

We also wanted to analyse if there were any difference between the sentiments during price hike vs. price dip days. We found that influencers' tweets are generally very positive even on price dip days as seen in the figure below.

	Overall compound score (Price Dip)	Overall compound score (Price Hike)
Vitalik Buterin	0.99	0.96
Elon Musk	0.77	0.99
Nayib Bukele	0.82	0.92
Changpeng Zhao	-.10	0.99
Bloomberg	0.99	0.98
Goldman Sachs	0.80	0.81

Figure 10: Compound Score Comparison between Price Hike and Price Dip Days

To evaluate if there is a correlation between the sentiment of the general population based on their post/tweets from Reddit and Twitter, and if the sentiments could be used to predict price change of Bitcoin, a linear regression model was used. The results are seen in the figures below. R-score for twitter general population (rise/drop), twitter general population (hike/dip), reddit general population (rise/drop), and twitter general population (hike/dip) were 0.015, 0.103, 0.001, and 0.041 respectively.

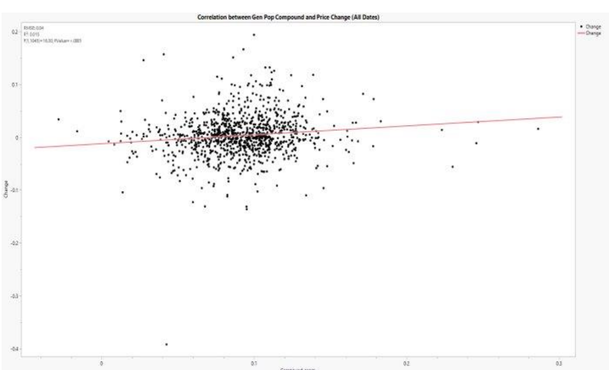


Figure 11: Correlation of Twitter Genpop vs Price Rise/Drop

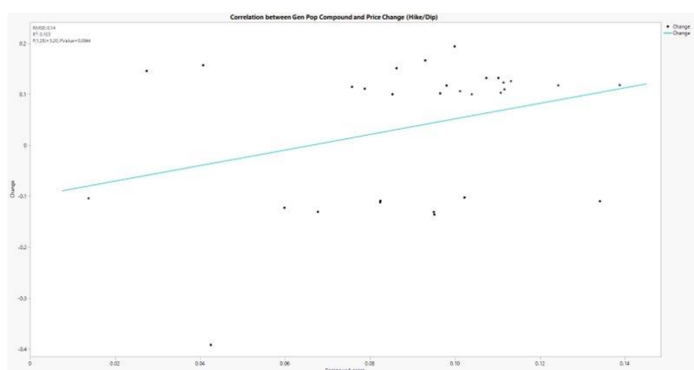


Figure 12: Correlation of Twitter Genpop vs Price Hike/Dip

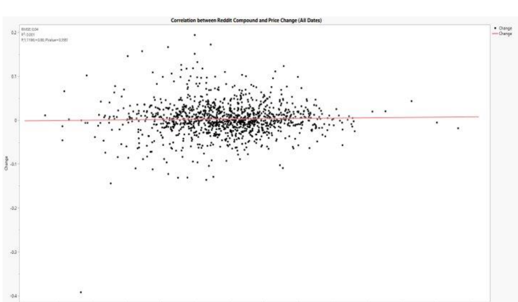


Figure 13: Correlation of Reddit Genpop vs Price Rise/Drop

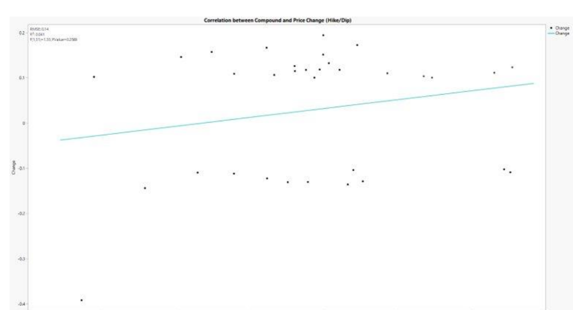


Figure 14: Correlation of Reddit Genpop vs Price Hike/Dip

5.3 Topic Modelling

The topics uncovered during Price Hike and Dip days are largely similar with topics related to Coronavirus, the 2020 United States' Presidential Election, Markets and China. This is expected as these topics dominated mainstream media from year 2019 to 2021. However, several unique topics during days with significant decrease in price were uncovered.

- Unique topics on price dip days: "Tariffs", "Xi Jin Ping", "Recession", "Fight", "Trade", "Huawei"

Figure 15: Topic Modelling for price hike days (left) and price dip days (right).

With the discovery of unique topics during days with significant price decrease, it can be concluded that these were the aggravating factors being discussed on social media, that contributed to the decrease in prices of cryptocurrency. We can infer that in late 2019 and 2020, China-United States trade war was at its height, and market fears spread to crypto-markets which caused the large sell-off and the drop in cryptocurrency prices.

6. Discussion of Findings

6.1 Social Influence on Twitter

Through PageRank score calculation, we can see that high number of followers does not equate to a high importance score, using follower counts as weightage. This is because the importance of a Twitter account is dependent on the quality of the followers (i.e. followers with high follower count).

This is particularly important for analysing Twitter subjects because it is widely known that Twitter contains fake accounts with very few followers, but high follower count, used to inflate one's account follower count. (Adetunji, 2022). Similar to the concept of page farming, it is important we consider the quality of followers in Twitter.

6.2 Sentiment Analysis

Based on the confusion matrix analysis, we can see that there is better accuracy and F1-score between sentiments of crypto-related tweets by influencers/institutions, to explain large changes in Bitcoin prices, as compared to non-significant Bitcoin prices changes. Hence, the +/-10% approach is a more effective filter for describing large price changes that may be attributed to the influencer's sentiments on social media.

There are also biases found in influencer's tweets as we can see that their tweets' compound scores are generally positive even during Price Dip days, which could be attributed due to their possible vested interest in Bitcoin/cryptocurrencies, hence they would still want to tweet positively to try and encourage followers to invest in the coin despite price dips. Therefore, investors should not depend on so-called "crypto experts" in attempt to predict price changes.

Due to the low R-squared value of the linear regression between price change and general population sentiment compound scores of both Twitter and Reddit, we can see that social media sentiments have low correlation with the behaviour of price changes. Hence, it can be concluded that the prediction of Bitcoin price change would not be possible by just analysing the sentiments of the general population.

6.3 Topic Modelling

Through Topic Modelling, we found that the prices of cryptocurrencies are affected by macroeconomic factors geopolitics, and general market sentiment. Issues between countries and major economies can add uncertainty in terms of policy direction (i.e. banning of crypto farming) or amount of investments in the alternate investment space. This was cemented in the discovery of top words of topics in days with significant price decrease related to the year 2019 – 2020 China-United States trade war and recession.

This is an interesting finding as the prices of cryptocurrencies are mainly determined only by how much the next person is willing to pay, and it is not grounded in fundamental finance. However, when the economy faces turmoil and uncertainty, the prices of cryptocurrencies are affected accordingly. For a new investor entering the cryptocurrency market, this is an important point to note.

6.4 Conclusion & Future Work

To conclude, our analysis shows that there is no causal relationship between the influence of an influencer/institutions and their follower count. Topics shared on their social media platform might a factor on their influential score. Lastly, no one should attempt to predict price change of bitcoin solely by looking at a one-dimension variable (sentiment of users). After our analysis, multiple improvements were identified for possible future studies to improve on the results of our findings:

- **Building a Cryptocurrency Sentiment Analyser.** Through our research, we identified that the in-built VADER sentiment analyser from the NLTK library was not able to comprehend cryptocurrencies' context (to the moon) and acronyms (HODL) and would provide a neutral sentiment score to this text. To improve the accuracy of the results, a specific sentiment analyser for cryptocurrency should be trained by using labelled tweets and reddit posts.

To conclude, our analysis shows that there is no causal relationship between the influence of an influencer/institutions and their follower count. Topics shared on their social media platform might a factor on their influential score. Lastly, no one should attempt to predict price change of bitcoin solely by looking at a one-dimension variable (sentiment of users). After our analysis, multiple improvements were identified for possible future studies to improve on the results of our findings:

1. Building a Cryptocurrency Sentiment Analyser. Through our research, we identified that the in-built VADER sentiment analyser from the NLTK library was not able to comprehend cryptocurrencies' context (to the moon) and acronyms (HODL) and would provide a neutral sentiment score to this text. To improve the accuracy of the results, a specific sentiment analyser for cryptocurrency should be trained by using labelled tweets and reddit posts.

2. Correlation of Bitcoin and S&P500. Since there is a significant correlation between Bitcoin and S&P 500 (Fulton, 2022), influencers/institutions such as Jim Cramer or Fed Reserves that possibly have an influence over the S&P500 should be selected for the research to provide different dimension to the analysis.

7. Contributions

Person	Contribution
Jordan Ong Zhi Rong	• Planning, Sentiment Analysis, Report, Slides, Pre-processing Tweets/Reddit posts, Code compilation
Kenneth Low Yan Wei	• Planning, Generating Ideas, Report, Slides, Sentiment Analysis, Topic modelling
Peace Tay Jiunn Ching	• Planning, Generating Ideas, Report, Slides, Scrapping of Follower IDs (Twitter), Network Visualisation, PageRank Centrality
Tan Yu Yan, Rachel	• Planning, Generating Ideas, Report, Slides, Confusion Matrix, Correlation Graph
Widya Tantiya Yutika	• Planning, Report, Slides, Scrapping of Influencers' and Institutions' Tweets, Scrapping of Reddit Posts and Sentiment Analysis on Reddit posts.
Yap Pin Yaw	• Planning, Generating Ideas, Report, Slides, Influence Score, Language Detect

8. References

Adetunji, J. (23 May, 2022). *How many bots are on Twitter?* Retrieved from theconversation.com: https://theconversation.com/how-many-bots-are-on-twitter-the-question-is-difficult-to-answer-and-misses-the-point-183425#:~:text=Twitter%20reports%20that%20fewer%20than,Officer%20Parag%20Agrawal's%20public%20response.

Chipolina, S. (25 Sep, 2020). *Why do Altcoins Usually Follow Bitcoin?* Retrieved from Decrypt: https://decrypt.co/42834/why-do-altcoins-usually-follow-bitcoin

Fulton, R. (14 June, 2022). *Is Bitcoin Correlated to the S&P 500?* Retrieved from We are relentlessly reimagining the world of tomorrow — one that is built on innovative technology, fueled by market-moving insights and driven by forward thinking.: https://www.nasdaq.com/articles/is-bitcoin-correlated-to-the-sp-500

The Investopedia Team. (21 March, 2020). *How Blockchain is Revolutionising Content Distribution*. Retrieved from Investopedia: https://www.investopedia.com/news/how-blockchain-revolutionizing-content-distribution/