University of Reading

Department of Computer Science

Analysing the Sentiment of Twitter Data on the Bitcoin using LSTM

Ching Yat, Cheng
eh801690@student.reading.ac.uk

Supervisor: Miguel Sanchez Razo

A report submitted in partial fulfilment of the requirements of
the University of Reading for the degree of
Master of Science in Data Science and Advanced Computing

16th July 2022

Declaration

I, Ching Yat Cheng, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Ching Yat Cheng
16th July 2022

Abstract

Cryptocurrencies are rising into one of the mainstream exchange mediums in different areas like the verification of asset transferring and the transaction of finances. Among that, Bitcoin is the main phenomenon from the beginning. Compared with the traditional financial tools, there has not been enough solution that can integrate sentiment analysis based on computational linguistics and Natural Language Processing (NLP) to learning algorithms for predicting the price of cryptocurrencies. On the other hand, the influence of social media is becoming significantly important in that market. The predictive power of Twitter shows in different fields of events, especially the market related to finance. In this paper, the LSTM algorithm of Bitcoin price prediction with the sentiment of Twitter data had been introduced. The sentiment scores were extracted by the public option from the different Twitter users. Furthermore, the Root mean squared error (RMSE) was applied for comparing the learning result between the net market data and the market data with sentiment for understanding the ability of Twitter sentiment data. As the result, the project proved that there is the presence of correlation between the Twitter sentiment data and market price. However, the insufficient amount of Twitter sentiment data had not shown any improvement in the prediction of cryptocurrency price.

Index Terms — machine learning, price prediction, cryptocurrency, social media, sentiment analysis.

1. Introduction

The most rapid-growth digital currency nowadays is a cryptocurrency, formed by powerful cryptography algorithms. Also, crypto is both an alternative form of a traditional exchange medium and a virtual accounting system. The related cryptography is used to protect the financial transaction, verify the asset transfer, and sustain a certain number of extra units. The decentralized system design is developed based on blockchain technology, which adopts a distributed ledger synchronized across multiple computer networks [1]. Bitcoin is the most well-known e-currencies, which is the first cryptocurrency created by the pseudonym Satoshi Nakamoto in 2009. Bitcoin was also published as a public, open-source design to prevent any organization or institution to take control of it [2]. Since the date Bitcoin was released so far, there are approximately 20,000 cryptocurrencies that use similar technology in the market. However, the whole market capitalization of a decentralized currency is $1.025 trillion in 2022, whereas Bitcoin values the highest current market cap that over $0.42 trillion [3], which is over 40% of the capitalization of the entire market.

The cryptocurrency market has drawn attention significantly due to the incredibly raising in market value. Some customers regard it as an actual exchange medium for transactions, whereas the rest of them treat it as an opportunity for investment. Caused by many retail investors and several investment organizations entering the market, the result of cryptocurrencies prices had extremely fluctuated. For example, the value of Bitcoin raised nearly 11 times from $4875.51 on 12 March 2020 to $57617.03 on 21 February 2021. Nevertheless, the price had decreased by approximately 40% in just only 3 months [4].

Unlike traditional currencies, the mighty huge volatility of cryptocurrency is causing uncertainties about whether someone regards cryptocurrency as a real currency or a financial investor. Due to the difference between traditional currencies and cryptocurrency, it cannot directly apply similar methods from another financial market to estimate how the volatility happens. Therefore, predicting the future price of the cryptocurrency market has become a new challenge for both financial investors and researchers. For increasing the accuracy of cryptocurrency prices, various sources of heterogeneous datasets are needed such as online blog content, newspaper articles, discussion forum comments, and social media posts. Among these, the social media platform called Twitter showed an incredible influence and it affected the price of cryptocurrencies massively. One of the best-known examples is that Elon Musk shared the tweet about DOGE Coin in May 2021, and the related crypto price was sent up to 30% immediately after Musk's tweet [5]. Nevertheless, the use of Twitter data requires filtering unnecessary information and needs to analyse the sentiment with an appropriate tool for obtaining the most accurate cryptocurrency price prediction. In addition, due to the complexity of the recurrent neural network, all different models have their specific usage. Therefore, selecting the machine learning algorithm and structuring the model for predicting the price of crypto are other challenges.

On the other hand, Sentiment analysis had been expected that it can be applied into a great deal of financial market [6][7]. At the same time, a lot of evidence shows that Twitter data from the influential figures maybe affect the price of the cryptocurrency directly. However, there are lack of experiment that attempt to discover the direct relation between sentiment analysis and cryptocurrency market.

Therefore, three main objectives that corresponding to the before-mentioned problems. First of all, by studying other research, plenty of valuable paper shows the possible methods or theories that can apply to this project. Thus, the first objective of this paper aims to refer to some previous studies to find an effective way to support this research. Next, due to the lack of research to investigate the cryptocurrency with sentiment analysis on Tweets data. Therefore, the research aims to discover the possibility of improving the accuracy of Bitcoin prices with sentiment data. Thirdly, there are some historical cases that popped up in the news and said that Twitter data would affect the price of the cryptocurrency. Hence, a researcher attempted to analyse the sentiment for exploring the actual impact of Twitter in the cryptocurrency market.

For achieve the objective of this paper, the Twitter sentiment data analysis embedded into the prediction model with a recurrent neural network had been developed in this project. the three before-mentioned objectives could be dealt with by several related tools and methods. First, the LSTM is a time-based recursive neural network that can perform a better prediction accuracy than other neural network models for processing and handling time-series data. Second, the proposed model applied the method called Valence Aware Dictionary and sEntiment Reasoner (VADER), which can analyse the content from tweets and generate a sentiment score by natural language processing. Lastly, the historical Bitcoin data and pre-processed sentiment from tweets content had integrated and modulated into the final-tuned machine learning model.

This paper's remainders are organized in the following order. Section 2 discusses the related work suggested in the literature that was proposed by previous research. Section 3 involves the methodology for the experiment implementation. Section 4 evaluates the project results, findings from the results, and the limitation of the project. Finally, Section 5 summarizes the project, reflects the criticality of the solution, and examinates any possible improvement.

2. Related Work

A state-of-the-art related crypto prediction and sentiment analysis reviewed in this section. There are mainly 3 categories classified from related work: (I) the perspective of sentiment from economic psychology and financial; (II) Prediction of the stock market by statistical model and machine learning model; (III) the prediction model related to the crypto market data.

2.1. Economic Psychology

Kahneman et al. have framed the 'Prospect Theory' which proposes that the decision of finance is not just only the value itself, but remarkably affected by emotions and risk. [6] Another similar work in the psychology of economic decisions supported that idea. Lucey et al. [7] proposed that the alterations of feelings that are broadly experienced by the majority would affect how people make their decision for investment. Thus, some patterns can be predicted in the market price. Based on their insights, social media data is expected to be useful for the financial market with some techniques like sentiment analysis would be possible to discover patterns that can influence the market.

Taking into consideration the existence and universalness of media, particularly social media networks, the next step of the work inquired into how it influences the sentiment of users and hence economic market trading. Huang et al. in [8] explained how optimistic emotions from social media caused a vigorous influence on the return of stock of different fields' industries, and unexpected extreme sentiment forecasts a high volume of trading. Furthermore, Yesha Mehta and Yogesh Chandrakant Funde [9] found that consumers inform purchasing decisions by referencing comments and opinions on social media platforms, based on their convenience and effectiveness. The rising number of researchers referred to these perceptions and aimed to discover the relationship between sentiment data from social media platforms and economic market trading.

2.2. Machine Learning with sentiment in Stock Market

Joshi et al. in [10] retrieved and took unquantifiable data like the articles related to economic news to analyse how sentiment data affect financial markets. The Dictionary-based approach model had been developed for analysing the sentiment of financial news. The models that the paper proposed had achieved 80% accuracy in all their prediction models about the trend of short-term stock value movement, with algorithms training the dataset only containing the sentiment from news articles. On the other hand, Chang et al. [11] attempted to predict the future return based on the message board comments sentiment. This paper collected the sentiment that relied on a specific company's topic as a feature for forecasting the

future stock price, instead of the overall emotions from articles related to the whole market on the website. Compared with the method of using historical market data only, the approach with the sentiment analysis of message board comments showed a better correlation in small stocks, but not in large-capitalization stocks. Moreover, the prediction power of this approach showed that sentiments were more effective in the first 2-3 days but haven't enough evidence to prove that it functioned in long-term prediction. Correspondingly, Mao et al. in [12] extracted company-related tweets for discovering the correlation between Twitter posts and the Standard & Poor 500 (S&P 500). The result validates that there is a certain relationship between the price and the number of tweets in specific stocks in different sectors. Singh et al. in [13] applied Twitter sentiment data as an input to 5 different machine learning models for predicting the value of AAPL stock prices and the DOW Jones Industrial average index. The result in Support vector machines (SVM) showed a better performance which is 82% accuracy.

2.3. Machine Learning in Cryptocurrency Market

Due to the wide-ranging investment in cryptocurrencies in recent years. Therefore, the researchers employed different machine learning algorithms or models in this area, and they had also been driven to attempt to forecast the price variations in this market. Fekry et al. in [14] applied RNN and Long Short-Term Memory (LSTM) models to predict a Bitcoin price in the dataset of 15, 30, and 60 days. The preliminary result demonstrated that LSTM had better performance and less mean absolute percentage error (MAPE), which is 97.7% more accurate than RNN in 30 days dataset. On the other side, Devavrat Shah and Kang Zhang in [15] proposed a method that applied Bayesian regression for predicting the variation of Bitcoin price, and the result illustrated that their investment had been gained 2 times more within a period of 60 days. Each research both shows evidence that machine learning, statistical modelling techniques, and neural network could predict the price or trend of cryptocurrencies. However, the topic of sentiment with neural networks in the crypto market is still a relatively new topic, there is not much research applying social media or news data to the neural network for forecasting the historical price of crypto.

2.4. Conclusion

Established on the experiment and discovery from the research outlined above, the sentiment data could benefit to forecast of the financial market, and the machine learning model could be expected to improve the prediction of the crypto market price. Furthermore, the LSTM model had been proven in [14] that could have a preferable performance to RNN. Therefore, this project aims to build up a LSTM prediction model of Bitcoin price in several keyways. Differentiated from the article [10], Twitter data had been used for analysing sentiment instead of news articles. Finally, in place of predicting a return from the investment profile like [11] and [15], the project intends to predict the actual value of the Bitcoin price. Therefore, RMSE is used for measuring the average difference between the predicted price and the actual price of Bitcoin, and it would better estimate how accurate the prediction model forecasts.

3. Methodology

The thesis achieved to retrieve, extract and analyse the Bitcoin price with the use of sentiment analysis and a recurrent neural network. In the following table, the chapter of a methodology section will be shown.

Table 1: the chapter of a methodology section

Chapter 3.1	System Design
Chapter 3.2	Data Collection
Chapter 3.3	Pre-processing
Chapter 3.4	Sentiment Analysis
Chapter 3.5	LSTM network modelling
Chapter 3.6	Summary

3.1. System Design

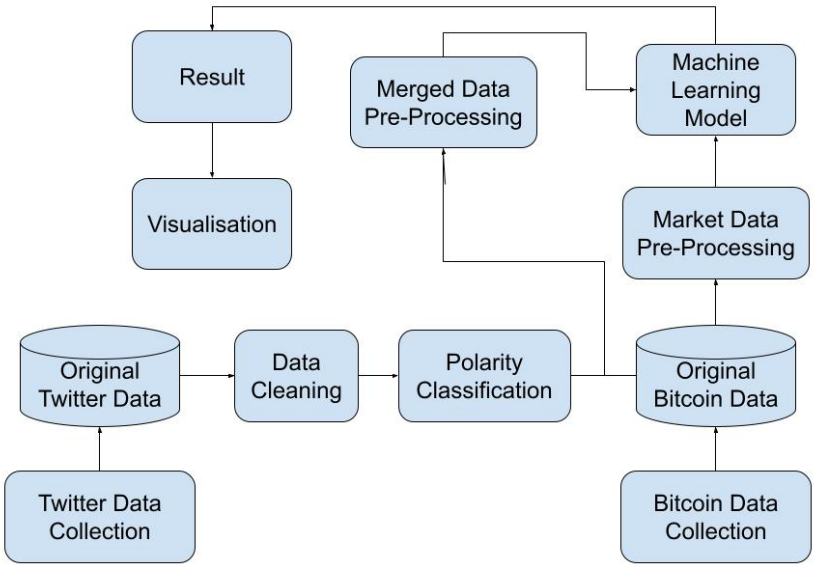

Figure 1 Project Architecture

The design of the LSTM prediction model with sentiment analysis had been divided into an end-to-end architecture which shows in Figure 1. Firstly, plenty of research shows that sentiment analysis would affect the result of the recurrent neural network model to predict the financial market. Therefore, the sentiment data had been extracted from the related Tweets, which were collected by Twitter API. Also, the historical Bitcoin market data will be collected by Kaggle. Secondly, the tweets data cannot be direct merged with the Bitcoin market data. Thus, the collected tweets' textual data is required for pre-processing such as deleting all unnecessary data, removing the duplicated content, and cleaning the stop word from the sentences. Thirdly, the social media data had been classified as sentiment polarity for analysing the sentiment score, and the calculated sentiment data had been saved independently with date and time. Fourthly, the time of the collected Bitcoin market data is a timestamp format instead of a date and time format, which is not

matched the sentiment data. Hence, the pre-processing of the historical data is needed. Fifthly, the LSTM model has required to set up the level of layers and the number of nodes. Therefore, the LSTM model in this part had been adjusted for obtaining better results. Sixthly, Twitter sentiment data is embedded in the original Bitcoin market data that is matched by time and date, and both merged data and original data are used as input for the prepared machine learning model. Seventhly, the prediction models' output had been plotted for visualizing the result. Finally, the result of both two had been compared for discovering the effectiveness of the sentiment data.

3.2. Data Collection

For feeding the data into the model of this project, there are two primary data that had been collected by web scraping, which are the Bitcoin-related tweets and Bitcoin market data (minutely). The selected period of the collected Bitcoin market data and Twitter data is from 1 July 2022 to 30 July 2022.

- Twitter data: Twitter API had been applied to extract tweet content by hashtags. For example, the symbol of Bitcoin is BTC, therefore, the tag #BTC was used for searching the related tweets on Twitter. To collect the data within a certain period, the academic research account from the Twitter developer portal had been applied. Afterward, there are two more related functions that can be used. One of the Twitter API's premium functions is called "search_30_days", which just permits a user to collect a maximum of 250 requests and 25000 tweets per month. Based on the limitation of that function, the project collected the data for exactly the month of July 2022. Finally, the number of extracted tweets is 29762 and the size of the data is approximately 4 MB. The originally extracted tweets data contained username, number of followers, number of friends, tweet content, length of the tweet, tweets ID, date, time, lang, number of likes, and number of retweets. The code of extraction shows in Appendix A.

	User	Tweets	Length	Tweet ID	Date	Time	Lang	Like	Retweet
0	souverwins	LOSSLESSNFT now this has my attention FFDefend...	140	1542674091798192128	2022-07-01	00:00	en	0	0
1	__itsmeeddie	El Salvador bought today BTC at each Bitcoin i...	140	1542674089227259904	2022-07-01	00:00	en	0	0
2	DarthDire	You re a bro ETH is a dumpster fire It clogs u...	140	1542674083015544832	2022-07-01	00:00	en	1	0
3	GraceMartino10	busy you are far blockchain btc eth eth Solana...	139	1542674081849389056	2022-07-01	00:00	en	1	0
4	Fomoxuser	NFTGIVEAWAY for x Spook PHASE NFT Follow xyz L...	144	1542674080347742208	2022-07-01	00:00	en	0	0
...	...	...	...	...	...	...	...	...	...
21044	_Ninjasclp_	My first long in months and you want to counte...	112	1553530540799803393	2022-07-30	23:00	en	0	0
21045	Hannah__vo	In a bear market one should remain sensible an...	140	1553530536236285954	2022-07-30	23:00	en	8	0
21046	rickustrades_	trades I am looking to get a better read on BT...	140	1553530531706445824	2022-07-30	23:00	en	0	0
21047	cixcix72	k holders CertiK done Listing on Lbank Mexc so...	140	1553530527629680641	2022-07-30	23:00	en	0	0
21048	Jommy9191	Get support Crypto Deposit Contribution via th...	140	1553530514610434049	2022-07-30	23:00	en	1	0

Figure 2 Sample Twitter Data

- Historical Bitcoin market data: Originally, the dataset of minutely Bitcoin historical data was planned to extract by the Yahoo! Finance API. However, the API function for getting market data minutely had been deleted by them. To solve this problem, the alternative dataset had been collected from the Kaggle dataset

called "400+ cryptocurrency pairs at 1-minute resolution" [17]. the collected data related to BTC-USD is from 2013-04-01 to 2022-08-02, which covered the selected period of this project. The data contain some information such as time, open, close, high, low, and volume. Among these, all the prices inside the dataset had been represented by USD dollars. Also, the format of representing the time is a timestamp format instead of a date-time format.

Dates	Time	open	close	high	low	volume
2013-04-01	00:07:00	93.25	93.30	93.30	93.25	93.300000
2013-04-01	00:08:00	100.00	100.00	100.00	100.00	93.300000
2013-04-01	00:09:00	93.30	93.30	93.30	93.30	33.676862
2013-04-01	00:11:00	93.35	93.47	93.47	93.35	20.000000
2013-04-01	00:12:00	93.47	93.47	93.47	93.47	2.021627
...	...	...	...	...	...	...
2022-08-02	21:21:00	22969.00	22970.00	22970.00	22969.00	0.000277
2022-08-02	21:22:00	22970.00	22956.00	22970.00	22956.00	0.004560
2022-08-02	21:23:00	22956.00	22956.00	22956.00	22954.00	0.106921
2022-08-02	21:24:00	22955.00	22953.00	22955.00	22946.00	0.110245
2022-08-02	21:25:00	22956.00	22967.00	22967.00	22956.00	0.209465

Figure 3 Sample Bitcoin Data

3.3. Pre-processing

The original extracted tweets dataset had 29762 tweets that were collected hourly within the period of time that had been mentioned above. However, the content of every tweet is not written in English. Therefore, the first step of pre-processing is to delete all the non-English content tweets. As the result, the remained dataset is just only 21049 tweets. The next step of the tweet pre-processing is to delete all the noisy content from the tweet since social media contents are relatively more casual than traditional media like news articles. Normally, social media content might be formed by the combined expression including URLs, emojis, symbols, URLs, tags, and mentioning other users. Furthermore, Twitter provided different unique features like user referencing, like, and retweets. The data relating to these features should be correctly extracted. Thus, the raw tweets dataset is required to be restructured to create clean content that can be more suitable for the sentiment analyser. An extensive number of data cleaning pre-processing had been implemented for standardizing the tweet content and reducing the unrelated data scale. The general processing had applied such as the following steps before pre-processing to reduce the meaningless content:

- Alternate all contents to lowercase
- Change the dots that are more than 1 to a space
- Displace the spaces that more than 1 to a space
- Erase all the quote symbols (“or ‘)

After the step of general processing, there are several pre-processing steps required to fully clean the data. The example of the steps had been shown in Table 2. The code of the cleaning steps had been shown in Appendix B.

Table 2 Pre-processing step of tweet example

	Steps	Result
1	Raw Tweet	rt @documentingbtc https://t.co/evo4oickk1 this clip of @aantonop speaking in 2013, now has over 1 million views andreas first ever public talk about #bitcoin.
2	Delete “RT” in front	@documentingbtc https://t.co/evo4oickk1 this clip of @aantonop speaking in 2013, now has over 1 million views andreas first ever public talk about #bitcoin.
3	Delete URLs, and mentions	this clip of speaking in 2013, now has over 1 million views andreas first ever public talk about #bitcoin.
4	Delete number	this clip of speaking in, now has over million views andreas first ever public talk about #bitcoin.
5	Delete Hashtags	this clip of speaking in, now has over million views andreas first ever public talk about .

As the paper mentioned before, the raw dataset of historical Bitcoin market information extracted from Kaggle required some pre-processing due to the time format and the range of date. First, the original time data is a milliseconds timestamp. By converting timestamp to date-time format, the variable type of the time column needs to change from String to Date-Time. After that, the converted column splits into Date and Time separately. Once obtaining the Date column, the range can be selected by "pandas.loc". Also, Both Date and Time can be used for matching the sentiment data.

3.4. Sentiment Analysis

In the part of sentiment analysis, the document level of every tweet had been focused on. All tweets inside the dataset had been treated as one document, which can generate a single row of sentiment scores. Generally, sentiment analysis for the textual content had mainly 2 approaches such as rule-based analysis and machine learning-based analysis. A machine learning-based approach is a supervised approach intrinsically that demands a sufficiently large enough dataset for learning the different characteristics from the dictionary of text. For analysing the textual content, machine learning approaches apply a classification method, which can classify the text. On the other hand, A rule-based approach is a training-free approach that can analyse textual content without machine learning modelling. A set of rules is the result that can classification into 3 polarities such as positive, neutral, and negative. Also, it applies the dictionary for determining the sentiment of words and combining the sentiment score of every word in the sentence to generate a sentiment score. The final sentiment score can estimate how positive or negative the sentence is. A rule-based approach is also known as a lexicon-based approach or dictionary-based approach since

the rule is come from lexicons. The main direction of the project objective is to develop a recursive neural network with Tweet sentiment data for discovering the effectiveness of the social media data and exploring the probabilities of increasing the accuracy in the prediction of Bitcoin, instead of innovating in the process of natural language processing. Therefore, the project applied a rule-based approach for sentiment analysis instead of a machine learning-based approach.

VADER approach had been applied in the project for the measurement of the sentiment from every tweet content in the collected dataset, which is the tool based on lexicon-based or rule-based to analyse the sentiment from the textual content. Also, the tool is expressly customized for analysing the textual content from social media platforms that contain particular emotional expressions. Once feeding the textual content into the VADER, the tool automatically generates 3 sentiment scores in the array such as positive, neutral, and negative. At the same time, the tool also collects the sentiment score from those three columns, adjusts the weight of the sentiment based on the rules, and calculates the compound score by summarizing the sentiment scores of every word in the sentence. Finally, the normalized sentiment score which is in the range of -1 to +1 will be the final result (-1 means extremely negative and +1 means extremely positive). Based on the step the paper mentioned before, the output of VADER is the compounded score that had been weighted and normalized. The outputted score provided one-dimensional data as a metric for measuring the sentiment in each tweet instead of the multi-dimensional of words or numerical data. Hence, the compound score that was outputted by VADER had been applied as the sentiment score of this project.

However, the Twitter data is not like the news articles content that only contains a list of words, but also involves a list of important data such as the number of friends, the number of followers, the number of likes, and the number of retweets. For example, famous figures must have a greater number of followers, which intuitively means their influential power is larger than the rest of the users, and they may have a big chance to impact the cryptocurrency market more effectively. On the other hand, the sentiment score extracted by VADER is not considering any of them except the textual content itself. Therefore, the final metric for analysing the sentiment is not just the output of VADER but also measured the side information from Twitter, such as the number of likes, retweets, friends, and followers. The finalized sentiment score combined all the elements by the following equation:

$$
\begin{aligned}
Sentiment = {} & VADER\ Compound \\
& * (1 + Normalized\ Likes \\
& + Normalized\ Retweets \\
& + Normalized \left(\frac{Followers}{Friends}\right))
\end{aligned}
$$

In the equation, the number of followers had been divided by the number of friends to prevent the tweets from bots had been counted into the machine learning model, since plenty of robots create meaningless tweets every day. The robot accounts will follow a huge number of users but only a minority of users will follow them back. Therefore, after weighing the number of followers by the number of friends, the effectiveness of the bot account will be decreased, or drop to nearly 0. On the other hand, the number of

likes and retweets is significant information that reflects the popularity of the tweet. Even though the celebrated tweet may have a chance that can affect the cryptocurrency market more, the sentiment from the users who are not famous is also important. Therefore, the +1 had been added to both the number of likes, retweets, and weighted followers, which means the maximum score of opinions from the famous figure is only 2 times more than the others. Also, all the extra information had been normalized for reducing the scale of the final output. The score weighted by the equation shows the single row of the data. For matching the sentiment data to the historical Bitcoin market data, the final calculated score by the equation from different tweets within the same date and hour had added up as a feature to fit into the cryptocurrency data for the prediction of the price. The code of extraction shows in Appendix C.

	Date	Time	Final_Sentiment
0	2022-07-01	00:00	0.196672
1	2022-07-01	01:00	0.091800
2	2022-07-01	02:00	0.181696
3	2022-07-01	03:00	0.227096
4	2022-07-01	04:00	0.218911

Figure 4 Sample sentiment data

3.5. LSTM network modelling

The most significant component of this project is the recursion neural network model. The correlation between social media data and a influential metric like the price of the Bitcoin market in USD had been captured by the trained model result. Successfully generating the difference between the data with or without sentiment and confirming the truth of the correlation of them are the indispensable feature of the machine learning model. The design of the machine learning model will be discussed in the following.

Firstly, neural network learning is one of the important parts of machine learning. Multiple hidden layers form a recurrent neural network (RNN), which is a class of artificial neural networks that imitate the process of decision-making in the human brain. The cycle in RNN is formed by every connection in the middle of nodes, which authorizes output from several nodes to influence consecutive input to the matched nodes. However, the simple RNN may not has enough ability to handle the problem of the explosion of exponential weight or disappearing gradient with recursion, which let RNN hard to capture long-term temporal correlation [16].

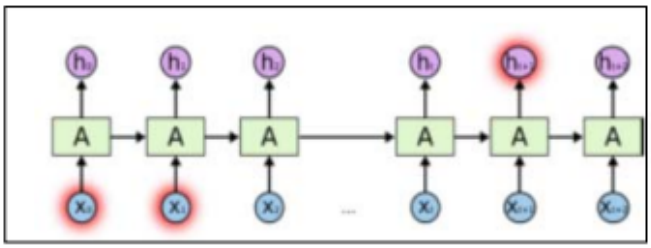

Figure 5. Detailed architecture of RNN

Long Short-Term Memory (LSTM) had been created at 1997 by Hochreiter & Schmidhuber [18], which is like the regular RNN but forms by different module type. Yet, many academic paper and research project built on top of the original LSTM and universalized the LSTM model. Like RNN, the LSTM recursion neural network had been assembled by recurrent consistency. Furthermore, RNN is an older version of LSTM. The largest disparity between RNN and LSTM is the connection among with the hidden layers of RNN. The detailed architecture of RNN had been showed on Figure 5. LSTM are sharing a similar architecture with RNN, but the memory cell of architecture in hidden layer shows a wide range of difference. The Figure 6 shows that LSTM had applied three particular gates that can beneficially resolve the problem related to gradient.

In addition, the disadvantage of RNN had been showed on Figure 5, which is the wide range of the information between input from X0 to Xt. Therefore, the end of the RNN is not able to get the relevant previous information like the X0, which cause the former memory becoming useless over the time since the fresh memory will cover it. Therefore, the combination of RNN and LSTM was applied to the project to solve these problems.

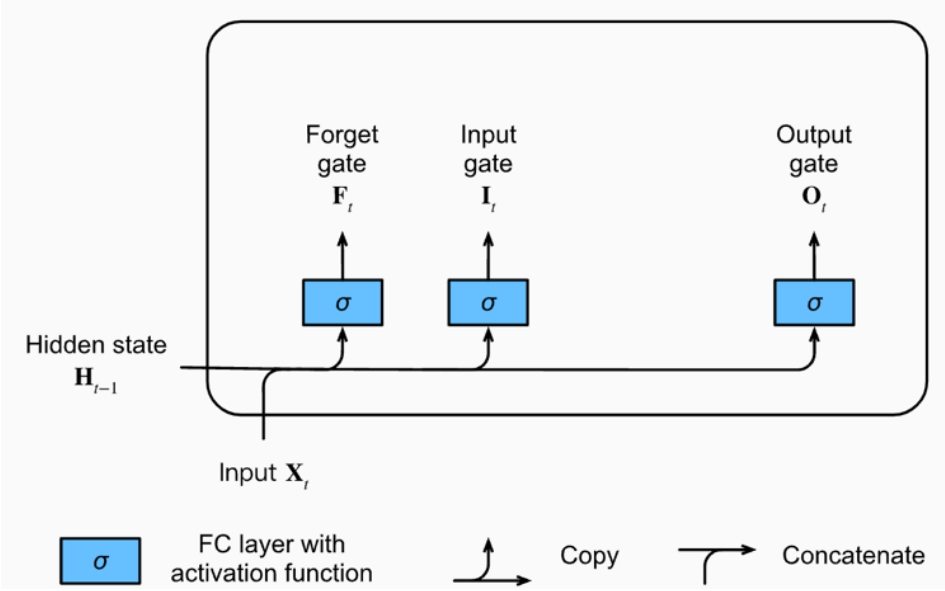

Figure 6. Gate in an LSTM model

Before fitting the data into the LSTM network model, there are required several steps to transform the data into a suitable format. First of all, the correlation of each column of data had been checked for avoiding the high similarity between x and y, which cause the unusually high accuracy of the result. As the correlation had been shown in Figure 8, “open”, “close”, “high”, and “low” show high similarity. Therefore, there is only a “close” between four of them had been selected as input.

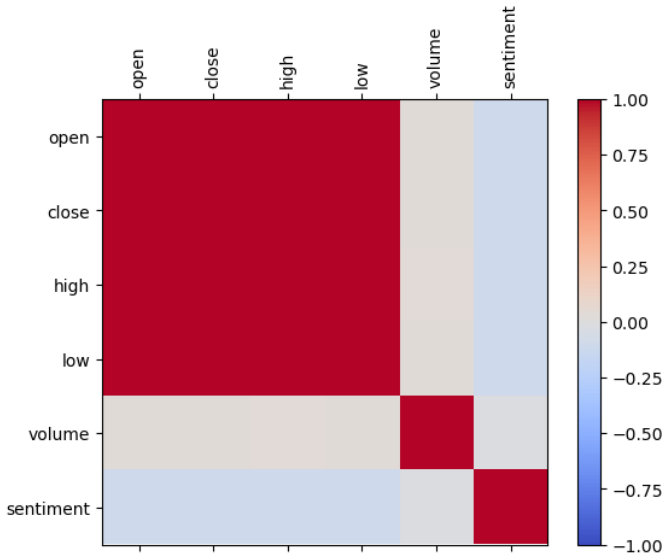

Figure 8. Correlation of the final dataset

Due to the wide gap in numbers between each column. Therefore, the MinMaxScaler() function from sklearn.preprocessing had been applied to normalize the dataset. Afterward, since the LSTM model requires the data in 3D array format. the self-defined function called series_to_supervised() had been applied to get a multi-variate time series and the return of the dataset had been framed as a dataset for supervised learning. Also, the expected result of the function is a time series Pandas DataFrame that is suitable for supervised learning, and the last 3 columns will be erased. The code of the function is in Appendix D and the result of the function is in Figure 9.

	var1(t-10)	var2(t-10)	var3(t-10)	var1(t-9)	var2(t-9)	var3(t-9)	var1(t-8)	var2(t-8)	var3(t-8)	var1(t-7)	...	var3(t-3)	var1(t-2)
10	0.258496	0.497923	0.002053	0.252489	0.497923	0.003758	0.262069	0.497923	0.001570	0.263646	...	0.011192	0.270340
11	0.252489	0.497923	0.003758	0.262069	0.497923	0.001570	0.263646	0.497923	0.008324	0.266392	...	0.000927	0.266294
12	0.262069	0.497923	0.001570	0.263646	0.497923	0.008324	0.266392	0.497923	0.002370	0.277034	...	0.004964	0.264161
13	0.263646	0.497923	0.008324	0.266392	0.497923	0.002370	0.277034	0.497923	0.008513	0.276519	...	0.004708	0.258990
14	0.266392	0.497923	0.002370	0.277034	0.497923	0.008513	0.276519	0.497923	0.003640	0.267615	...	0.003720	0.257611
...	...	...	...	...	...	...	...	...	...	...	...	...	...
43058	0.828871	0.513839	0.000907	0.827931	0.513839	0.000182	0.831960	0.513839	0.000098	0.827497	...	0.000969	0.835127
43059	0.827931	0.513839	0.000182	0.831960	0.513839	0.000098	0.827497	0.513839	0.000213	0.827841	...	0.000074	0.835050
43060	0.831960	0.513839	0.000098	0.827497	0.513839	0.000213	0.827841	0.513839	0.000076	0.827755	...	0.000399	0.831617
43061	0.827497	0.513839	0.000213	0.827841	0.513839	0.000076	0.827755	0.513839	0.000024	0.832132	...	0.000148	0.832670
43062	0.827841	0.513839	0.000076	0.827755	0.513839	0.000024	0.832132	0.513839	0.000047	0.836010	...	0.000250	0.832990

43053 rows × 33 columns

Figure 9. Sample output of series_to_supervised() function

However, after the process of the series_to_supervised() function is not still a 3D data structure. Therefore, the dataset requires restructuring by some configs. First, the variable called “n_min” is creating an n of lag for the observations, and n here is 10. The variable “n_features” is counting how many features the original dataset has, so here are 3. Before transforming the data to the 3D array, the data had been split into 66.6%

of training set and 33.3% test set. After the process in Appendix E, the shape of train X is (28000,10,3), train Y is (28000,), test X is (15053, 10, 3), and test Y is (15053,).

Python library "Keras" had been applied to the project experiment as a deep machine learning framework. For predicting the minutely Bitcoin closing price by using the previous N minutes, a deep neural network model compounded by 6 layers of LSTM structure had been designed. There are 2 Dropout layers had been structured between 3 layers of LSTM the value of the Dropout layers is 0.2, and all the number of nodes in different LSTM layers is the same. In the end, the output layer called "Dense" had built to export a certain number of the result. In the case of this project, the output is the predicted value of the Bitcoin price. Therefore, the number of Dense layer outputs is 1. The summary of the model setup is shown below:

Layer (type)	Output Shape	Param #
lstm_9 (LSTM)	(None, 10, 50)	10800
dropout_6 (Dropout)	(None, 10, 50)	0
lstm_10 (LSTM)	(None, 10, 50)	20200
dropout_7 (Dropout)	(None, 10, 50)	0
lstm_11 (LSTM)	(None, 50)	20200
dense_3 (Dense)	(None, 1)	51

Total params: 51,251
Trainable params: 51,251
Non-trainable params: 0

Figure 9. Summary of project LSTM model

The number of Epoch defaulted as 1000, but the model that we had setup allow earlier stop. The total number of Epoch in the model without sentiment is 129 and training time is 9189.86. The total number of Epoch in the model with sentiment is 219 and the training time is 11837.18.

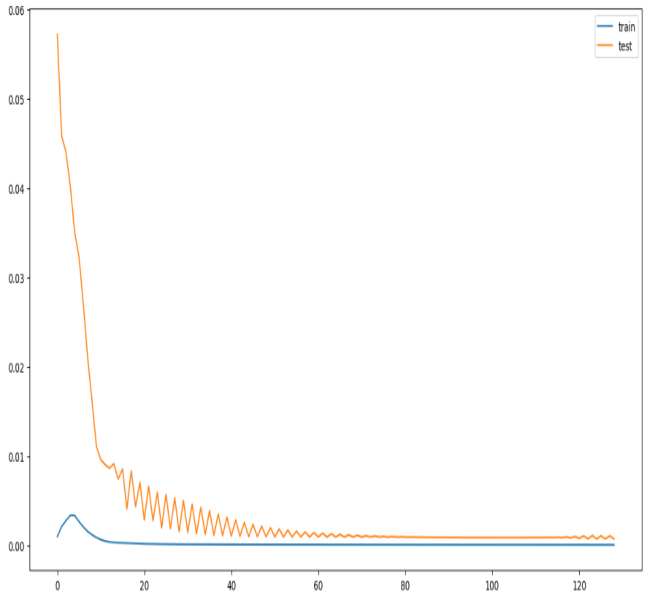

Figure 10. Plot MSE within Epoch without sentiment

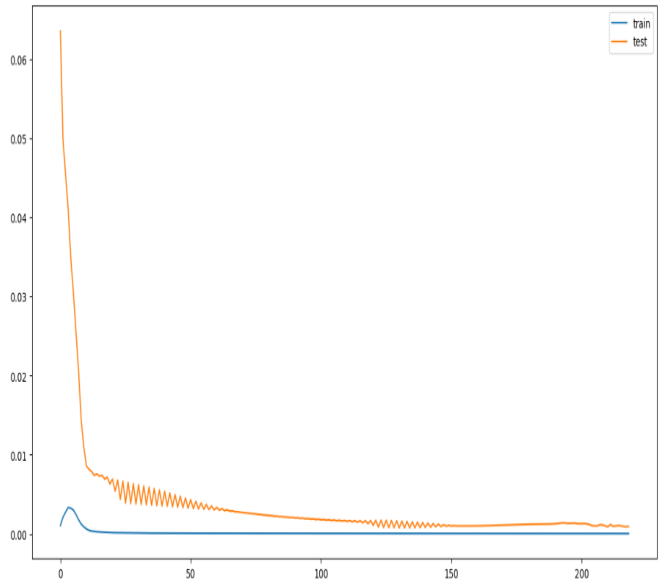

Figure 11. Plot MSE within Epoch with sentiment

3.6. Summary

In this session, the system design had been shown in chapter A, which included Figure 1 for displaying the system design flow chart. Also, every step that related to the project had been described clearly. In chapter B, the selected data collection period, the type of the data, and the method of data collection had been shown. In addition, the detailed number of rows of the dataset, the called function of API, and the data contained in the dataset had been described by the type of data separately. The sample results of the collected data were shown in Figures 2 and 3. Chapter C illustrated the Twitter textual content pre-processing one by one in Table 2 and showed the method of handling the timestamp data from the Bitcoin market dataset. In chapter D, the sentiment analysis approaches had briefly compared. Furthermore, the tool that was applied to the project called VADER had also been detailly described, and the designed equation of weighting the sentiment score was shown clearly. The sample dataset was also shown in Figure 4. Finally, chapter E explained the difference between RNN and LSTM in Figures 5 and 6. Also, the self-designed LSTM model had been described and construed clearly including the data restructure and split Training-Testing set.

4. Result

The previous research had only shown the results from the machine learning model consisted of the generated sentiment scores from news articles or forum discussions for predicting the historical stock trading data. However, there are still insufficient data that can prove that social media data can also be suitable for the cryptocurrency market. On the other hand, compared with social media posts, the new articles show less performance in the short term. the social media platform that the project selected as a sentiment input is Twitter, which shows the information immediately. Nevertheless, social media platforms normally allow users to release any content except criminal content. Therefore, a lot of misinformation had been posted by different users or bots. For these reasons, the project would like to discover the actual relation between sentiment data and historical Bitcoin trading data. Through the series of the process from data collection to machine learning model development, the 2 most significant results have been generated for showing the discovery of the project.

4.1 The correlation between sentiment data and Bitcoin price

	close	sentiment	volume
close	1.000000	-0.104398	0.029502
sentiment	-0.104398	1.000000	-0.020195
volume	0.029502	-0.020195	1.000000

Figure 12. Correlation Table of the final dataset

The correlation table for showing the relationship between each data shows in Figure 12. Firstly, the dataset had been reduced to only 3 columns which are close price, sentiment, and volume. By the table

that Figure 12 shown, the value of correlation between sentiment and close price is -0.1044, which is not strong but still have a certain relationship.

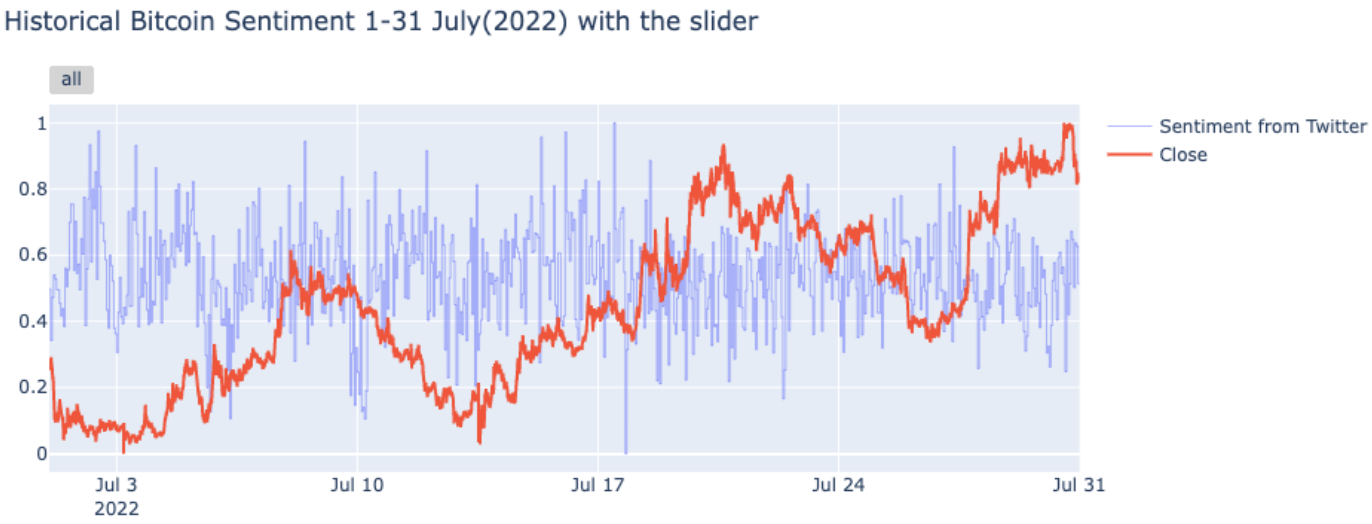

Figure 13. Line Graph of historical data and sentiment with time series

The Price of closing Bitcoin price and sentiment data had been captured in Figure 13. The minutely Twitter sentiment data and the matched bitcoin price had been plotted in the graph. the figure also shows some of the peaks of the emotional score correlated to the previous hour of the closing price. Furthermore, due to the same volume of tweet data every hour, the number of tweets does not affect the sentiment calculation. Finally, even though the graph shows the correlation of the peaks between sentiment data and the closing Bitcoin trading price, the sentiment data is relatively more exaggerated, which means it is oversensitive.

4.2 The prediction of Bitcoin price

Figure 14. Graph between actual and predict value with sentiment

Figure 15. Graph between actual and predict value without sentiment

Table 3. The result measured by RMSE

	Training with sentiment	Testing with sentiment	Training without sentiment	Testing with sentiment
RMSE	**213.07**	**178.72**	**202.52**	**153.31**

The Root Mean Square Error (RMSE) had been selected as a measurement method to survey the prediction result of the project. Compared to Mean Absolute Error (MAE), the RMSE metric assesses the model's prediction power in continuous value, which shows a better performance than MSE. Furthermore, the unit of RMSE is matched with the unit of the data variable that is expected to predict, which means the output is showing the difference between the prediction value and the actual value. The RMSE metric is measuring the difference between two values, which means that if the RMSE value is smaller and the model performance is better. Based on Table 3, the smaller result is the testing set without sentiment data. Also, the performance of both the training set and testing set with sentiment shows worse than the dataset without sentiment. By comparing case by case, the value of the training set with sentiment is 213.07 and the value of the training set without sentiment is 202.52, the error between them is 10.55. On the other hand, the value of the testing set with sentiment is 178.72 and the value of the testing set without sentiment is 153.31, the error between them is 25.41.

As the average closing price value is over 20000 during the period, the errors between the 2 models are extremely small. The graphs in Figure 14 and Figure 15 plots the actual value as blue line, training set predict value as green, and testing set predict value as red. The figures are showing that the shape of the prediction between the 2 models is basically the same. However, the sentiment data that was previous collected is just creating a "white noise" to influencing the prediction model without any benefit.

4.3 Summary

In this session, a certain correlation had been proved in Figure 12 and Figure 13. For the information provided in Figure 12, the correlation between Twitter sentiment data and Bitcoin market closing price was

negative, which means if sentiment data is higher and Bitcoin closing price is lower. However, Figure 13 showed that some of the peaks between sentiment data and market data are matched, but the significant problem of the sentiment data was the value goes more exaggerated than the actual market data and caused the data less meaning. On the other hand, the results of prediction models illustrated that the dataset with sentiment data caused a worst prediction than the dataset without sentiment data. The estimated problem of lower accuracy is because the sentiment data is too sensitive and it causes the exaggerated value, which leads to affecting the model decision like white noise.

5. Discussion and Analysis

The project developed 2 LSTM neural network models, the model without sentiment and the model with sentiment analysis for predicting the price of Bitcoin. By comparing the models of the market data with sentiment and the market data with the sentiment, the sentiment data shows irrelevant to increase the accuracy of predicting the daily price of Bitcoin. Precisely, the 2 models show similar predictive results in both the training set and test set, but the sentiment data is supposed to complete the model and increase the accuracy of the prediction result. However, a correlation between the market data and the Twitter sentiment data had been found that is exist, which means that the sentiment data is not completely useless. Therefore, the estimated of the possible problems in the sentiment data had been concluded in the Finding chapter.

5.1 Findings

The Twitter sentiment data is one of the main elements of the project since the sentiment data from social media had been confirmed by plenty of academic research papers and application projects that can be beneficial to predict the stock market value. However, after training and analysing the dataset, the collected sentiment was proved that it may not benefit too much from the prediction accuracy with LSTM neural network even though the correction had been shown on the correlation table. After deep consideration, there are 2 possible issues had been concluded.

Firstly, the number of tweets that had been collected during the data collection session is 40 per hour. Based on the Central Limit Theorem (CLT) [20], the findings had been warranted surely against if the sample size is over 30, which means it can be represented the population. Nevertheless, the case of sentiment analysis in the project is time series data, and every minute in time series data is not exactly the same as the previous. Therefore, the amount of data and the period of the collection can be increased for extracting more meaningful data. Secondly, the VADER model can only extract the mean from a single word instead of treating the sentence as complete data. Therefore, the compounded sentiment scores are only based on the weighted score from every single word. The meaning of a word inside a sentence can be different than the meaning of a word itself. Therefore, the original meaning of the sentence may be misinterpreted. Even though the data that was scrapped from Twitter is relatively less, some of the

correlations still exist. However, the opportunity for improvement in the prediction of the cryptocurrency cannot be proven in the project.

5.2 Limitations

There is plenty of limitation in the data collection, which is restricted by the API limitation and the cost of upgrading the API. First, the Twitter API “search_tweets” function is not allowed users to send too many requests to extract the data within 15 minutes, and every request of the search function is only allowed for 100 tweets. Furthermore, only the data within 7 recent days permit extraction. Therefore, another function from Twitter API called “search_30_days” had been selected. However, this function also exists a lot of limitations. The search_30_day function is a premium function that can only allow enterprise and academic users to access it. Even though the permit had been successfully proved by the developer platform admin, the free academic version is not allowed to collect the data easily. The restrictions of the free version called “Sandbox” can only retrieve the date within 30 days, and a monthly quota of requests is 250 and a monthly quota of tweets is 25000. The developer portal provided a button for upgrading the package from Sandbox to Premium, but the price of upgrading the service cost up to $2499 USD, which is extremely expensive. Therefore, based on the limitation, the tweet can only extract hourly and 40 tweets for every hour.

5.3 Summary

In this session, the result of the project had been mentioned in the first part, which discussed the result of the project had no significant difference but the correlation between sentiment data and market data existed. Therefore, the possible issues of the dataset collected from Twitter and the discovery of the project had been listed in the chapter of Finding. In the first place, due to the limitation of the Twitter API, the sample size of the tweet is not tight and large enough, and the time series data cannot directly apply a statistical theory called CLT because of the data type difference. In the second place, the VADER is a word-based sentiment analysis tool instead of a sentence-based sentiment analysis tool, which means the meaning of a sentence may be misinterpreted by the tool. Finally, the limitation of Twitter API had been detailly explained such as the request and tweet extract limitation, the account restriction etc.

Package	Sandbox	Premium
Time frame ⓘ	Last 30 days	Last 30 days
Tweets per request ⓘ	100	500
Counts vs. data ⓘ	Data only	Both
Query length ⓘ	256 characters	1024 characters
Operator Availability ⓘ	Standard	Premium
Rate limit per minute ⓘ	30 requests/min	60 requests/min
Enrichments ⓘ	n/a	URLs, Polls, Profile Geo
Dev environments ⓘ	1	2
Monthly Tweet cap ⓘ	25K	5M
Monthly Tweet cap ⓘ	25K	5M
Rate limit per second ⓘ	10 requests/sec	10 requests/sec

Figure 16. Twitter API limitation between Sandbox and Premium

6. Conclusions and Future Work

Bitcoin is the main phenomenon of the cryptocurrency market which is a relatively new exchange medium or financial tool. Sentiment analysis of Twitter data had been proven that can beneficial for predicting the stock market. Therefore, the project aims to discover the probabilities of increasing Bitcoin prediction accuracy by sentiment analysis and discover the correlation between Twitter sentiment data and Bitcoin price. During the project, data on historical Bitcoin prices and Bitcoin-related tweets were collected. After several data cleaning processes, the tweet data can be fit into the sentiment analysis tool for extracting the sentiment scores. Finally, the sentiment data and market data merged into a single dataset fed into the LSTM recursion neural network model. Also, the data containing only market data is fed into the same model to compare the result for observing the performance of sentiment data. As the result, the project outcome shows that the data without sentiment data had a better performance, the collected sentiment data proved that is not useful due to the insufficient amount of data, but a correlation among them had been discovered. For proving the estimated issues of the prediction model, there are 3 possible methods that can be tested in the future:

- Bidirectional Encoder Representations from Transformers (BERT): It is a sentence-level sentiment analysis tool, which can break down a sentence into a lot of short sentences or mask some words inside the sentence to predict the actual meaning in the sentence. Furthermore, BERT can also predict the sentence next to this sentence, which can also understand the relationship between two sentences. By applying BERT, sentiment analysis can generate a sentiment score more precisely.

- Real-time extraction and prediction: Based on the limitation of Twitter API that can only collect the data within 7 days without quota. Therefore, the real-time prediction model can resolve the problem of the shortcoming of insufficient data. The real-time model can be set up by the own server or cloud computing, but both methods require relatively high computational power, which means a high-cost budget is needed. However, it is a possible method that can solve the problem, therefore, the approach will be tested in future work.

- Reinforcement Learning: It is an agent-based learning machine learning algorithm that is different than supervised and unsupervised learning. reinforcement learning requires the agent to decide their actions instead of supervised learning which requires pre-defined labels in the data. Based on the theory of reinforcement learning, it can be assumed that the input data size can be lesser. In the case of this project, one of the probable problems lead to the model does not beneficial from the sentiment data can be fixed if the reinforcement learning had been applied with LSTM.

7. Reflection