ANALYSIS OF EXTERNAL FACTORS INFLUENCING BITCOIN EXCHANGES

Project Progress

THE UNIVERSITY OF SYDNEY

Information Technology Capstone Project

COMP5703/5707/5708

Group Members

1. Chenxiao Cai (470158649)
2. Xingchen Zhou (460122838)
3. Chuhan Wang (312062133)
4. Sihong Huang (450000469)

ABSTRACT

The goal of this project is to investigate and analyze what are the external factors that influence bitcoin price exchanges and develop a deep-learning model to predict bitcoin next-day closing price base on the outcomes of factor analysis. The previous work does not take a wide range of external factors into consideration and merely achieves sub-optimal prediction performance. In addition, it does not provide an in-depth news sentiment analysis. The task is achieved by applying statistical analysis on bitcoin's supply and demand side factors, hard commodities factors and USD/EUR exchange rate. Relevant news is another category of factors and they are processed by sentiments analysis. The prediction modelling part presents Support Vector Regression (SVR), Linear Regression and Long Short-Term Memory (LSTM) with different combination of features as inputs to predict next-day bitcoin price. The results of the experiments show that LSTM outperforms the other two models with 115.17 RMSE and 62.1% accuracy of next-day bitcoin price (rise/drop) classification.

Keywords: bitcoin, sentiment analysis, feature selection, linear regression, SVR, LSTM

CONTENTS

1 INTRODUCTION

Bitcoin (Nakamoto, 2008) is the world's most valuable cryptocurrency. Bitcoin is traded on more than 40 exchanges worldwide, accepting more than 30 different currencies. According to blockchain.infor(2019), its current market capitalization is 9 billion, with more than 250,000 transactions occurring every day. As a currency, Bitcoin offers a new price prediction opportunity because it is relatively young and generates volatility far greater than the fiat currency (M. Brie‘re and Szafarz, 2013). In terms of its openness, it is also unique in terms of traditional fiat money; there is no complete data on the legal transaction of cash transactions or currency.

As an investment, The benefits and risks of Bitcoin investment coexist, professional organizations have different attitudes toward Bitcoin. Forbes said bitcoin is the best investment of 2013, In 2014, Bloomberg said bitcoin is one of the worst investments of the year(Hill, Kashmir. 2015). In 2015, bitcoin topped Bloomberg's currency tables (Steverman, Ben, 2015). The price trend of Bitcoin is a topic that thousands of investors and technicians want to study and analyse. In our project, our aim is to develop a physical model that utilizes machine learning and deep learning techniques to accurately predict bitcoin price movements over a given time period by analysing the effects of external factors. By using our model, the investor can know how the external factors influence bitcoin price, thus faciliates their investment decisions.

The purpose of this paper is to study and analyse the external factors that affect the price of bitcoin and use machine learning as well as deep learning algorithm to predict bitcoin price. In this project, we will use data visualization and analysis techniques to retrieve, store, and analyze data from one or more cryptocurrency exchanges. We will also visit websites, read news, articles in order to discuss and analyze external factors that impact cryptographic price exchanges. We will then use two machine learning techniques and one deep learning technique to predict price movements and to complete their forecasts based on real-time exchange of data.

2 LITERATURE REVIEW

2.1 Literature review introduction

This section summarizes past work related to various avenues of bitcoin, including bitcoin bascis, factor selection and analysis, modelling. The selection criteria is mainly about the keyword such as bitcoin, machine learning. The form of the article contains conference paper, journal paper or research report thesis. Through the summary analysis of the literature, determine our project direction and evaluate project performance. At the end of this section, the limitation of past work and project will also be mentioned.

In order to identify external factors impacting bitcoin price movement, studying relevant journal article and related project is of great value for us at early stage because literature review is always a starting point that not only sparks our inspiration to conduct further research but also provides solid academic evidence to back up our argument. Analyzing external factors impacting bitcoin price movement is

vitally important since bitcoin price movement is highly erratic and unpredictable. However, driven by speculative investment behavior, investors are lured to invest in this cryptocurrency to enrich their investment portfolio. Recent price fluctuation leaves a lesson that one can never get lucrative return without fully understanding the insights of external factors that can potentially determine the future price movement. Therefore, if the problem can be solved by taking the most determining external factors into consideration and building a solid model to accurately capture the highly-volatile price movement pattern, the risk of bitcoin investment can be mitigated to an acceptable level. In terms of the scope, emphasis is put on journal articles and conference proceedings that either discussing the potential factors that may impact bitcoin price or predicting bitcoin price movement by utilizing machine learning models. These two kinds of literatures are highly related to our aims and objectives of the project, therefore obtaining valuable information and insights from these sources are essential to the successful delivery of the project. Lastly two waves of literature review are scheduled with different keywords. The first wave is a board research with keywords such as "bitcoin price factors". After collecting a certain number of articles and understanding the findings from the preliminary literature review, the second wave is more targeted and specific, with ad hoc search criteria on a narrow range of keywords such as "bitcoin price and gold price" and "bitcoin price and exchange rate." Most literatures are obtained from USYD library by using the library search engine. Thus, the source of our reference is highly professional and reliable.

2.2 Cryptocurrencies bitcoin

For this project, it is important to understand the principles of Bitcoin and the price changes of Bitcoin. The reading is summarized as follows: A cryptocurrency is a group of digital currencies that are not regulated by either party and use encryption as proof of work. Proof of work is an economic term for measuring the effectiveness of a service. In this article, solving the encryption function is the most important part of the cryptocurrency (Nian and Chuen, 2015). One of the most and oldest cryptocurrency that has been circulating around the world is bitcoin. bitcoin originated from a peer to peer electronic payment scheme proposed by Satoshi Nakamoto . For every transaction with bitcoin, a process must be done to solve a hash function as proof of work, This process is called mining,

Mining is done by either individuals or groups, and Bitcoin is offered to them as a reward, which is one of the most effective ways to get Bitcoin (et al, 2016). After acquiring bitcoin, people can trade on trading platforms at a certain exchange rate with US dollar, which is the main activity of Bitcoin transactions. Today, most investors use bitcoin investment as an effective mean to develop their wealth.

According to(Dwyer, 2015), based on historical bitcoin data from 2010 to 2013, the average return on bitcoin investment was 7.14, and the highest rate of return was 136.72 (**?**), but the lowest value also reached -41.78. This high volatility is consistent with the characteristics of high-risk, high-return investment forms. Therefore, we intend to analyze the external factors that affect the price of Bitcoin through mechanical learning, so as to better predict the price of Bitcoin and give investors a reliable investment advice.

2.3 External Factors Identification

2.3.1 Supply and Demand

Bitcoin are defined as either alternative method of payment, an investment (asset) or a commodity by different governments in the globe(Mandjee, 2015). Supply and demand is a significant economic factor that influences the price of financial instruments in the financial market (Lewis, 2017). Thus, investigation of relationships between bitcoin price and supply and demand variables, and the magnitude of impact of supply and demand are indispensable parts in this research. Bitcoin has limited supply, in total, 21 million bitcoins are expected to be realised, which contained by its hard-coded protocol. To date, miners have mined more than 80 percent of the total coin and the maximum stock of BitCoins will not change until 2040. Current bitcoin supply is inflating at around 4 percent annually and this rate will drop dramatically in 2020. Base on empirical analysis (Gronwald, 2015)states that bitcoin in circulation and its growth rate are known with certainty, so there is no uncertainty on the supply-side of bitcoin. Extreme price movements are the results of some speculative activities driven by demand-side changes. While (Ciaian et al., 2014) hold a different view regarding supply-side factors. Their econometric models show an increase in the stock of bitcoin leads to a drop in its price. However, the demand side variables such as unique bitcoin address used per day and days destroyed for any given transaction have stronger effect on bitcoin price than supply side variables. Moreover, trading volume and volatility seem to play important roles on determing bitcoin price both in short run and long run, verified by (Sovbetov, 2018). (Hayes, 2015) also supports supply-side factors may have impact on bitcoin price. Regression models show that computational power employed in mining for units of bitcoin and the rate of unit coin production (per minute) does have influence on bitcoin price exchange. Moreover,in (S. Vassiliadis, 2017),the authors show that bitcoin price is affected by the transaction volume and the transaction costs, and there is a strong correlation between them. In addition, the price of gold and oil also have a significant correlation with bitcoin price. However, the correlation between stock and virtual currencies is weak. Finally, in the economic research proposed in ((M. Balcilar, E. Bouri, R. Gupta, D. Roubaud, Can volume predict Bitcoin returns and volatility? A quantiles-based approach, Econ. Modell. 64 (2017) 74–81.), the trading volume and bitcoin income are analyzed and found a causal relationship between rate and volatility. They concluded that volumn of transaction can be used to predict bitcoin price, but can not predict the volatility.

2.3.2 Macroeconomic and financial indicator

Introduced in 2008, bitcoin has embraced a significant surge of investment interest from investors in the globe. Just like other investment such as stock, gold and financial derivatives, bitcoin price is said to be intrinsically related to the global macroeconomic and financial market condition. With reference to (Sukamulja and Sikora, 2018), macroeconomic indicator such as Dow Jones Industrial Average (DJIA) and gold price affect bitcoin price both in the short run and long run. They derived the results by applying Vector Error Correction Model in the short term and long term. The result reveals that DJIA and gold price are statistically significant in relation to bitcoin price. Specifically, both indicators are negatively

related to bitcoin price. Similarly, Vassiliadis, Papadopoulos, Rangoussi, Konieczny and Gralewski (2017) utilize cross-correlation analysis to determine the relation between bitcoin price and major stock market indices as well as gold or oil prices respectively. They found that there is a strong relationship between bitcoin price and stock market indices such as SP 500 index, NASDAQ index and Dax index. In addition, such statistically-significant relationship also applies to gold price or oil price. Moreover, Saefong (2017) upholds the negative relation between bitcoin and gold price. Saefong reported that some analysts believe a rising bitcoin price is the culprit of gold price decline in December 2017. The argument was backed up by the research deliverables by Zwick and Syed (2019), who extracted data from 19th July 2010 to 31st December 2018 and adopted threshold regression model to investigate the relationship. They found that the relationship between gold price and bitcoin price is not linear in the long-run. While the impact of gold price was negative and weak before October 2017, the relationship became strong and positive since then. Bouri,Gupta,Lahiani and Shahbaz's work (2018) also echoes with Zwick and Syed's findings as they have found an asymmetric and non-linear relationship between bitcoin price and gold price. The fruition was acquired by mining corresponding data from 17th July 2010 to 2nd Feb 2017 and utilizing advanced autoregressive distributed lag (ARDL) models for statistical analysis.

2.3.3 Alternative cryptocurrency investment

In a large number of literature, different authors have analyzed the relationship between bitcoin and other cryptocurrencies. Most experts or scholars believe that bitcoin as the most famous cryptocurrency has guiding significance for other cryptocurrencies, a change in bitcoin price is highly likely to lead to a change in other cryptocurrencies. In this section, we will focus on the impact of other cryptocurrencies on bitcoin prices based on existing literature. Litecoin, an early bitcoin spin-off or altcoin, was introduced in October 2011. Many altcoins have been created since then (WIRED.2017), Litecoin is very similar to bitcoin, Litecoin developers said Litecoin is the "silver" to bitcoin's "gold." As of February 24, 2019, the market value of bitcoin was approximately 67 billion dollars, while the market value of Litecoin was 2.7 billion dollars(WIRED.2017). bitcoin has been the main source of cryptocurrency since 2009, but Litecoin and others have joined the crpytocurrecny market to enrich investment portfolio. Multi-party research shows that bitcoin has guiding significance for Litecoin, and the price change of Litecoin can also be a good predictor of bitcoin price changes. When investors find that the price of Litecoin has dropped or increased significantly, people will involuntarily consider the price trend of bitcoin.Ethereum is another kind of cryptocurrency, for those who want to understand and implement cryptocurrency mining, there are significant differences between bitcoin, Ethereum and Litecoin mining. All three coins show the potential to innovate in different ways. However, one thing is clear: they all seem to yield lucrative returns from long-term investments. For pricing, the fact of comparing bitcoin and Ethereum is very simple. The bitcoin price increased by about 1,000, while the Ethereum increased by about 10,000 (Mint Dice August 5, 2018) . Despite the higher overall price, the data suggests that bitcoin investment may not be as good as Ethereum. There is another potential cryptocurrency here that comes from the newly issued

cryptocurrency. In the paper, "The Impact of Tether Grants on Bitcoin," Dr. Wang Chunwei discuss the relationship between new cryptocurrency such as the issuance of tether (USDT) and bitcoin, the findings shows that tether grants may follow bitcoin decline. But the impact of tether grants on bitcoin returns do not have any statistically significance, so tether issuance cannot be an effective tool for predicting bitcoin prices (Wang Chunwei 2018).

2.3.4 Attractiveness

(Ciaian et al., 2014) use a modified version of Barro's model to investigate the relationship between bitcoin price and bitcoin attractiveness to investors. As a result, they find that there is a strong correlation between bitcoin price and attractiveness, namely wiki views, number of new members and number of new posts. The rationale behind is that the value of bitcoin is determined by investor perception. Bitcoin per se does not have any monetary value, but if bitcoin is attractive for investors for the sake of investment profit, it will appreciate. When investors are no longer interested in bitcoin, its value will depreciate. Therefore, we believe that attractiveness plays a significant role on determining bitcoin price trend. Since number of wiki views reflects the popularity of the topic related to bitcoin while number of new members and number of new posts demonstrate investors' enthusiasm and involvement in discussion about bitcoin.
Karalevicius, Degrande and Weerdt(2017) applied lexicon-based sentiment analysis on news and blog posts to analyze the relationship between social media sentiment and bitcoin price movement. A finance specific psycho-semantic dictionary is used as sentiment analyser to give sentiment score for each news collected. As a result, they found that with the assistance of social media, investors are able to predict bitcoin price movement in the short run. In addition, the market tends to overreact the sentiment that will automatically correct the price in the long run. So investors will never earn abnormal return purely based on social media sentiment.

2.3.5 Mining

With the growing number of users in Bitcoin, understanding the factors that influence its price exchange becomes more important (Hayes, 2017) and (Al Shehhi et al., 2014) found more than 50 percent users participate in cryptocurrency mining. The main topic in this part will focus on mining difficulty and mining cost. The mining difficulty measures the time required to mine a single unit of bitcoin on average. The mining process is the core of bitcoin system, it needs to discuss and analysis the mining difficulty and mining cost during the bitcoin mining. (Hayes, 2017) introduces bitcoin market is worth 70 billions which deals with 60 million transactions per day for now. From that, Li analyzed the important factors which affect bitcoin value. In addition, Li used regression analysis to get the results. As a result, he finds bitcoin value occurs at the margin, the price is affected by relative cost of production. Specifically, electricity consumption is the main cost of mining.
(Al Shehhi et al., 2014) investigate 8 kinds of cryptocurrency and external factors. They use questionnaire to collect data from user. The results show that crpytocurrecny mining is highly skill-demanding and there are many miner communities where mining expertise and experience can be shared.

(Li and Wang, 2017) investigate the impact of exchange rate,mining difficulty and mining cost in relation to bicoin price. They use the ARDL model to test the stationary and non-stationary time series. They found that both exchange rate and mining difficulty are important external factors. Exchange rate by comparison shows a higher level of significance over mining difficulty.

2.4 Modelling and Prediction

2.4.1 Works on Financial Data Forecasting

Long-term research on the prediction of mature financial markets such as the stock market is very popular(Kaastra and Boyd, 1996). Bitcoin presents an interesting parallel because it is a time series prediction problem in the short turn (White, 1988). Traditional time series prediction methods, such as the Holt-Winters exponential smoothing model, rely on linear assumptions and need to be decomposed into trends, so that seasonal and noisy data could be effective (Chatfield and Yar, 1988) This method is more suitable for predicting tasks such as sales that have seasonal patterns. Due to a lack of seasonality and high volatility in the bitcoin market, these methods ineffective for this task. Deep learning provides an viable technical solution to tackle complicated tasks because of its state-of-the-art performance in similar projects. Due to the temporal nature of bitcoin data, recurrent neural networks (RNN) and long-term short-term memory (LSTM) are superior to traditional multilayer perceptrons (MLP).

In (Wah B W, 2006), the Regression Neural Network (RNN) is used to predict the price of several stocks, namely Citigroup, International Business Machines (IBM) and ExxonMobil. Through the 10-day forecast, Citigroup, IBM and Exxon Mobil's MSE are 0.131, 0.182 and 0.226, respectively, and RNN can be seen as an effective model for price prediction.

In (Kshirsagar G, 2018), a predictive model combining data mining and ANN is used to predict the company's stock price. Their research shows that artificial neural networks deliver reliable stock market prediction, but there is great potential to improve the model performance.

The bitcoin price prediction can be considered similar to other financial time series forecast problems, such as foreign exchange and stock forecasting. Several research institutions have implemented MultiLayer Perceptron (MLP) for stock price forecasting (Adebiyi A A, 2012) . However, the MLP model only analyzes one observation at a time. Instead, the output of each layer in the Recurrent Neural Network (RNN) is stored in the context layer. In this sense, RNN has gained various memories compared to MLP. Another form of RNN is the Long Term Short Term Memory (LSTM) network. The mechanism differs because LSTM can choose which data to analyze and which data to discard based on the feature importance. In (Federal Reserve Board, 2015), found that in the time series prediction task, both RNN and LSTM achieve high performance.

2.4.2 Works on bitcoin Forecasting

As a virtual currency, bitcoin shares similar trading properties with other financial investments, so the analysis and modelling of alternative financial investment also has guidance and auxiliary functions for the study of bitcoin price. Many projects attempt to explore various predictive models to predict price.

The relevant literature are summarized as follows:

(Mallqui and Fernandes, 2019) implement Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Ensemble Algorithms (based on Recurrent Neural Networks and the K-meanings clustering) to predict bitcoin price movement direction, maximum, minimum and closing price by analyzing both internal and external factors. It turns out that SVM achieved the best results for all predictions at 59.45 percentage. (Phaladisailoed and Numnonda, 2018) use bitcoin transaction data from the bitstamp website and adopt four distinctive machine learning models for training and predicting, namely Theil-Sen Regression, Huber Regression, Long short-term memory (LSTM) and Gated Recurrent Unit (GRU). Accuracy is measured by Mean Squared Error (MSE) and R-Square(R2). The performance comparison among the four methods shows that it is GRU that stands out in terms of MSE and R2 (0.00002 and 0.992). (Radityo et al., 2017) focus on implementing four ANN models to predict bitcoin exchange rate to US dollar, namely backpropagation neural network (BPNN), genetic algorithm neural network (GANN), genetic algorithm backpropagation neural network (GABPNN), and neuroevolution of augmenting topologies (NEAT). Mean Absolute Percent Error (MAPE) and training time are chosen as performance metric to evaluate the performance of each model. The result shows that BPNN is the best model as it only takes 347 seconds to train the model while achieving 1.998 percent MAPE.

In (Y. Peng, 2018), the authors combined the GARCH model with SVR. It evaluates the model performance of bitcoin, Ethereum and other cryptocurrencies, and uses this method to analyze the prediction performance of traditional currencies such as the Euro, the British Pound and the Japanese Yen. (All of these are measured in US dollars.) The authors used this method to determine the factors that influence the price of Bitcoin. In his research, the errors of RMSE and MAE obtained from high frequency data are much lower than the low frequency data.

In (H. Jang, 2018), the author selected relevant attributes to conduct time series prediction. The prediction is performed by using a Bayesian neural network (BNN). The performance of BNN is compared to that of support vector regression (SVR) and linear models. The authors collected daily bitcoin data from September 11, 2011 to August 22, 2017. The price and volatility of bitcoin were analyzed and predicted through training and testing of this data. The result reveals that MAPE is 0.0198 and 0.6302 for each model respectively. Therefore, BNN and SVR can be used for bitcoin price analysis and prediction.

2.5 Limitation

After reading the literature, we conclude that the academic community is particularly lacking the use of machine learning algorithms to predict bitcoin prices. Most scholars or institutions develop models to predict bitcoin prices, but they are limited by the range of data resulting in low accuracy. For example, (I. Georgoula and Giaglis, 2015) implemented a potential source model to predict the price of bitcoin, noticed a accuracy of 54 in 24 days' data, and a accuracy of 62 in 7 days' data. In (Rechenthin, 2014) Google also used a similar method to predict the volume and the price. However, due to the small sample size, the research is not convincing, in another hand the news on social media usually has false or

exaggerated phenomena, which seriously affect the prediction results. Liquidity is quite limited within Bitcoin exchange. As a result, the market is at greater risk of manipulation. For this reason, our report will focus on the research about the impact of news, and we will use scientific methods to analyze the impact of news on bitcoin prices.

In (Navickas, 2018), the paper confirms that there is a correlation between sentiment change and price movement. However, the machine learning model fails achieve desirable performance. Low-level machine learning knowledge must be gathered to properly evaluate the models and achieve better results in next iterations.

Another limitation is the inferior data quality. In (Kharde and Sonawane, 2016), although the authors collected more than 5 million tweets, only 2.5 million tweets can be used because half of the tweets collect only contains hyperlinks,emojis, hashtag or even false and exaggerated statement, providing less value to get insightful sentiment.

Although Karalevicius, Degrande and Weerdt's work specializes in performing sentiment analysis to determine the relationship between social media sentiment and bitcoin price. The project lacks depth of analysis since there are a vast variety of news. Studying all the news merely may provide an overall view on how social media will impact bitcoin price but fails to divide news into several categories. News is a juicy field that worth exploring in greater detail. For example, news can be divided into regulation and policy, privacy and security, and the financial market. Studying the various avenues of news opens a door for people to fully understand external factors from legal, security and market aspect. All these factors are vitally important when forming bitcoin investment strategy.

In (Numnonda, 2018), the author only choose Open, Close, High, and Low as selected features, and it may not be enough to predict the bitcoin prices since various factors, such as the reactions from social media, policiesand laws introduced by the government to deal with digital currency can all contribute to the rise and fall of bitcoin price. Therefore, in order to provide a comprehensive and holistic analysis and develop a compelling model, a wide range of data including both fundamental factors and technical factors ought to be collected.

2.6 Literature review conclusion

Overall, supply and demand factors statistically significant to bitcoin price exchange. Especially, the bitcoin economy size – unique address used per day, trading volume and volatility may have strongest effects. However, bitcoin price variation contributed by supply/demand factors may absorbed by other factors such as attractiveness variables in more general specifications (Ciaian et al., 2014).

In addition, it can be inferred that macroeconomic and financial indicator is an integral part of bitcoin price dynamics. The reason being that bitcoin is perceived as a form of investment, which is alternative to other major investment such as stock and gold. As the logic develops, once the major financial market suffers downturn, investors are likely to invest their capital in cryptocurrency. As a result, the surging demand for bitcoin will drive the price up. Therefore, we intend to select gold price and oil price as

proxies for macroeconomic and financial indicator in our project. Moreover, other cryptocurrencies have a certain impact on bitcoin price changes, while Bitcoin has a particularly clear guiding significance for other cryptocurrencies. Investors' consideration of multiple cryptocurrencies is conducive to investment and can predict the trend of bitcoin prices better.

Furthermore, solid evidence has been found to link the subtle or unsubtle impact of bitcoin mining difficulty to bitcoin price movement. When it comes to technical aspect, RNN and ANN are particularly favored by most related projects given its high level of accuracy and efficiency as well as low MAPE.

Lastly, social media sentiment is of much value to analyze and drill down. Since social media covers a wide range of topics ranging from economic, finance, regulation, politics to privacy and security. So in this project, we also intend to mine different categories of news to study how news from each category will impact bitcoin price movement.

These findings provide us with intuitions about what algorithm and method is a good fit to predict bitcoin price. In brief, the findings summarized from the literature reviews can be valued references for the design and planning for the project. Through literature review, we have got a clear perspective of the fundamentals of bitcoin dynamics,obtain a full list of external factors we may wish to carry further research. Mostly importantly, based on the literature review we fully understand the limitations of past projects. These limitations motivate us to conduct a more in-depth research,analysis as well as model design and development in order to close the gap in the field.

3 RESEARCH/PROJECT PROBLEMS

3.1 Research/Project Aims and Objectives

The aim of this project is to identify, analyze the magnitude of impact of different factors influencing bitcoin price by exercising feature selection techniques, fundamental analysis. The project also plans to develop solid models which utilize machine learning and deep learning technique to accurately predict bitcoin price movement for a given period of time In order to achieve the goal of the project, several overriding objectives ought to be considered, they are listed below:

1. Conduct research and perform academic literature review from reputable sources.

2. Summarize findings and identify factors that potentially impact bitcoin price movement from the literatures.

3. Collect relevant data from reliable sources.

4. Consolidate all the data into a single spreadsheet and transform data from different sources into a consistent format.

5. Write programs to conduct statistical analysis and provide visualization to determine the relationship between candidate factors and bitcoin price movement.

6. Write programs to perform sentiment analysis on bitcoin related news and derive daily and monthly sentiment score. Study the relationship between bitcoin price and news sentiment.

7. Write programs to apply feature selection and feature extraction techniques to rank the important of each factor and select factors with the highest level of importance for machine learning and deep learning modelling.

8. Build up machine learning and deep learning model to perform training.

9. Use the trained models to predict price movement.

10. Evaluate performance and optimize the model.

11. Finalize the model and provide conclusion.

3.2 3.2 Research/Project Questions

In this project, the high-level question that we intend to echo is “What external factors impact the price of bitcoin?”, “Among these factors, to what extent does each factor contribute to the price movement?” and “Can we use a model to generalize the pattern of bitcoin price movement?”

3.3 Research/Project Scope

Our group will first conduct literature review and identify 16 external factors (gold price, oil price, wiki review, number of relevant tweets, number of active addresses, Litecoin price, mining difficulty, bitcoin volume, USD/EUO exchange ratio, number of bitcoin transaction, privacy and security news sentiment, regulation news sentiment, WSJ news sentiment, 3days moving average, 9 days moving average and 15 days moving average) influencing bitcoin price movement. Then all the relevant data will be collected, consolidated and pre-processed to be ready for analysis. Further, our group is going to perform feature selection and a range of techniques to analyze and select the most important factors. Lastly, the project is also concerned with using the most important factors to build Linear Regression, Support Vector Regression and Recurrence Neural Network model to predict bitcoin price.

4 METHODOLOGIES

4.1 Data Collection

Data are acquired by two approaches: the first one is to directly collect data from web resource and the second is to derive from existing data by referring formulas or rules.

4.1.1 Numerical Data

There are various type of platforms storing bitcoin historical price and its relevant indicators. It is necessary to check existing tools and available APIs to facilitate data extraction from various platforms. CoinMarket API is a simple and commonly used API to extract historical cryptocurrency data from CoinMarket. In this research, daily bitcoin and litecoin open, close, high and low prices, as well as volume are collected by this API. Records of daily bitcoin transactions, active addresses, mining difficulty and the number of relevant tweets are captured from BitInfoCharts by a dedicated API found online. For other relevant financial derivatives and indicators such as open, close, high and low prices of gold price and oil price, and the ratio of USD and EUR exchange are collected from ”www.macrotrends.net”. A Pageview

analysis tool (https://tools.wmflabs.org/pageviews) provides the number of daily views of “bitcoin” page in Wikipedia.

3-day, 9-day and 14-day simple moving average of bitcoin price are calculated base on the formula (where A is the average in period n and n is number of time periods):

$$SMA = \frac{A_1 + A_2 + ... + A_n}{n}$$

Numerical data are processed separately with news by Python. Gold and oil prices, USD/EUR exchange ratio do not list on the weekends, these missing values are filled with their closest prices/ratios in order to match data formation of bitcoin price. Moreover, weighted prices of bitcoin, litecoin, gold and oil were calculated by getting averages of their corresponding open, close, high and low prices to make following analysis easier.

4.1.2 News

News are gathered from bitcoin magazine and the Wall Street Journal (WSJ). Bitcoin magazine is a major social media press reporting bitcoin related issues since 2012, ranging from market trends, analysis, regulations to security and privacy. Since the project aims at studying the legal and security impact on bitcoin price movement, bitcoin magazine can be a reliable knowledge source for our team to collect regulation news as well as privacy and security news. Due to the technical issues, our team failed to apply python scrapper to mine the news from bitcoin magazine. Instead the regulation and privacy news are collected manually. The news data is organized and consolidated in a single CSV document with date of publication, title, content, category and sentiment score. In total 605 articles are collected, including 96 privacy and security news, 16 scams news, 181 technical news, 100 laws and justice news and 212 regulation news.

The WSJ is one of the mainstream financial news presses in the globe and it prides itself being neutral, professional and authoritative in news reporting. Numerous readers visit the website and read WSJ news on a regular basis. Therefore, the WSJ is reliable source for news related to bitcoin market trend. In order to get bitcoin market news from the WSJ, python news scrapper is utilized. Given the fact that the University of Sydney has no membership on the WSJ, the python scrapper is only able to obtain date, category, title and summary of each news dating from 1st July 2015 to 23rd April 2019. However, it is strongly believed that in most cases title together with summary has already provided enough information about the polarity of the news. As a result, the python scrapper successfully collected 435 news. The news is organized and stored in a separate CSV file, the format is consistent with bitcoin magazine news.

4.1.3 Twitter Mining using Twitter API

Our group also attempted to utilize the Twitter API Figure 1 to mine bitcoin-related tweets. In this attempt, our group is only interested in obtaining popular tweets. We assume that a tweet is deemed to be popular when the number of retweets exceeds 100.

However, after conducting holistic scrutiny on the benefits and potential risks of selecting twitter as a

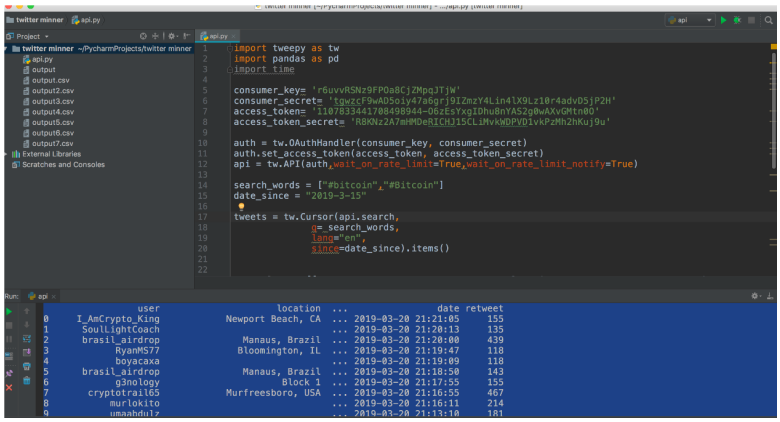


Figure 1. A screen shot of interface of the twitter API

feature, we decided to withdraw twitter from the candidate feature list for several reasons. First, the Twitter API has inherent limitations since it only allows us to mine twitter up to seven days. Our group failed to derive sufficient popular tweets about bitcoin from 2015 to 2019. Second, as we went through the twitter about bitcoin, we truly doubted the authority and reliability of twitter. Unlike news, everyone is free to express opinion regarding bitcoin on their own will regardless of what is honest and true. Therefore, twitter may not serve as a reliable knowledge source to study the bitcoin price movement.

4.2 Feature Selection

To investigate what are the external factors that influence bitcoin price exchanges and how bitcoin price is affected by these external factors, the first task to be completed in this project is to apply some statistical methods to show relationships among selected factors and bitcoin price in numerical way.

4.2.1 Correlation Coefficient

The correlation coefficient is a statistical measure gives value ranged between -1.0 to 1.0. It calculates the strength of the relationship between the relative movements of two variables. A positive value closer to 1 indicates a stronger positive relationship, whereas a negative value closer to -1 indicates a stronger negative relationship. A result of zero indicates no relationship at all. In this project, factors are evaluated by Pearson's correlation coefficient¡.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

Figure 2 indicates the technical factor - moving averages are highly correlated to bitcoin price since they are actually derived from bitcoin price by different time periods. Beside that, litecoin price, USE/EUR rate, number of relevant tweets and oil price are also highly correlated.

4.2.2 Tree-based Feature Importance Selection

Random Forest: is one of the mainstream machine learning algorithms. As a decision tree method, it can be used for tackling both classification and regression problem. Breiman (2001) defines random forest as

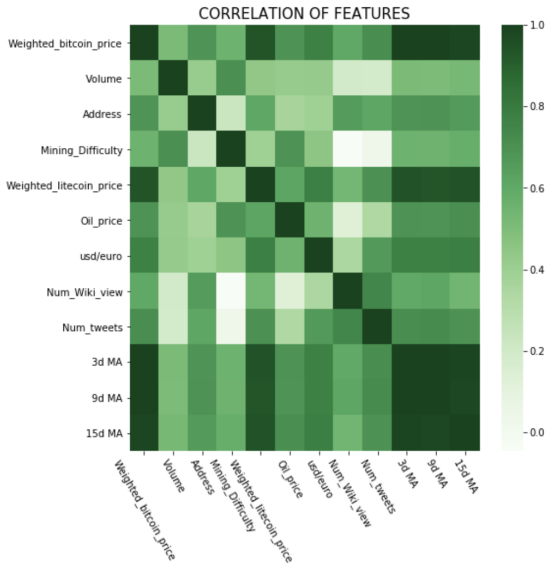


Figure 2. Correlation Coefficient

an "ensemble learning algorithm. Specifically, it comprises of two underlying algorithms, namely bagging algorithm and boosting algorithm. The basic random forest training and predicting process are briefly summarized below: 1. Set up a forest of trees and for each tree in the forest, utilize boosting algorithm to select training data and relevant features based on feature importance.

2. For each tree in the forest, an independent prediction will be performed given a test data.

3. The whole forest will nominee the most popular prediction from all tress in the forest. The nominated prediction then becomes the final output.

The flow chart demonstrates the basic workflow Figure 3 of random forest algorithm (only two trees in a forest):

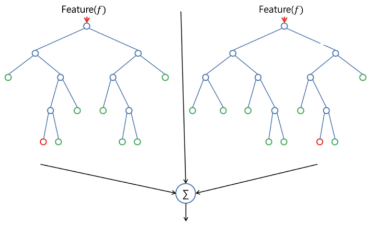


Figure 3. Random Forest Workflow (Niklas,2018 (Donges, 2018))

The reason why random forest algorithm is in our favor to conduct analysis is because random forest is able to provide a clear perspective of feature importance for all candidate factors. The feature importance ranking is expected to be a valued asset in our analysis for its interpretability in feature selection. Empirically speaking, result of feature importance is derived from the calculation of evaluation function.

The evaluation function considers incremental information gain for each candidate factor. The higher the information gain obtained from a factor, the more important the factor will be. Similar to information gain, level of impurity is also considered when determining feature importance, the feature importance increases as level of impurity falls.

In addition, random forest is a robust and efficient algorithm, the application of random forest algorithm covers a vast variety of industries, ranging from banking, finance, biotech to medical. The application of random forest algorithm in these industries has been observed to achieve state-of-the-art performance.

$$I(A) = \phi\left[p(y = 1|A)\right]$$

XGBoost: (Chen and Guestrin, 2016) describe XGboost algorithm as a scalable end-to-end boosting system. XGBoost is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning.When comparing to random forest algorithm, XGboost selects a set of independent CART trees for prediction. The function is provided below, where k refers the number of CART trees, F refers all CART trees, f refers to a specific CART tree: [= $\phi(X_i) = \sum_K^{k=1} f_k(X_i), f_k \in F$

One of the benefits for using XGBoost for analysis is the embedded feature importance function. As a gradient boosting algorithm, it is straightforward to derive feature importance for each CART tree when constructing the model. For each CART tree, feature importance is calculated based on information gain or level of impurity. The overall feature importance aggregates the feature importance result from all CART trees within the model. Therefore, XGBoost can be a reliable tool for this project to find out a list of important features, facilitating the feature selection and feature engineering process.

4.2.3 Relative Weight Analyses

Relative Weight Analyses(RWA) is to partition explained variance among multiple features to better understand the role played by each feature in a regression analysis. Comparing with RWA, traditional statistics (e.g. correlations, standardized regression weights), are potentially faulty or misleading information when concerning variable importance, especially when the variables are correlated with one another(Tonidandel and LeBreton, 2015). Since there are massive factors that potentially affect bitcoin price exchanges, and they are very likely to influence bitcoin price as a whole. This project aims to understand how each feature contributes toward bitcoin price, therefore RWA is a necessary technique to be applied for feature analysis. RWA transfers correlated variables to a set of new predictors which are the maximally related to the original predictors but are orthogonal to one another (Tonidandel and LeBreton, 2015). Besides obtaining relative weights themselves, adding confidence intervals with the weights and calculate statistical significance is a good way for testing results. Scott and James (2014) developed a RWA-Web that allow users to easily obtain variable weights and corresponding confidence interval tests of significance, which is available at: http://relativeimportance.davidson.edu/.

4.2.4 Mutual Information Regression

Mutual information is one of the most commonly used feature selection method. It measures how a presence or absence of a feature impact the reliability of a classification or regression decision making. In other words, it refers to the dependence between two variables. The value of mutual information is a non-negative value. A variable with greater mutual information denotes higher dependency.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} log(\frac{p(x,y)}{p(x)p(y)})$$

Mutual information will be used as one of the feature selection methods for this project since it provides reliable criteria to determine the dependence between two variables. In this project, mutual information will be applied to measure the level of dependence between each feature and bitcoin price.

4.2.5 Sequential Forward Feature Selection

The sequential forward feature selection (SFFS) is a feature selection algorithm that deals with feature selection based on feature importance. The algorithm begins with a empty list of features. In each round, a new feature will be added into the list and the corresponding performance metrics can be yield. The feature selection loop will end until the algorithm finds an optimal set of features that maximize the model performance. In this study, we use random forest classifier to apply the SFFS algorithm.

The purpose of applying forward selection is to select the columns (factors) that reject the null hypothesis - there is no relationship between two measured phenomena. This technique supports to select the factors that have relationships with bitcoin price exchanges within a defined significance level . This is a reverse version of backward elimination method, which gives the same results. Steps of forward selection (Chapter 10 Variable Selection): 1. Start with no features/factors in the dataset. 2. For all features/factors not in the dataset, check their p-value if they are added to the model. Choose the one with lowest p-value less than crit (significance level).3. Continue until no new features/factors can be added. Ranking factors from smallest p-value to largest p-value to get their importance on impacting bitcoin price exchanges in decreasing order. For those factors who have p-value greater than 0.05 (features left over by forward selection), a conclusion that they have no sufficient effect on bitcoin price can be made.

4.2.6 Recursive Feature Elimination

The recursive feature elimination(RFE) algorithm is adapted to rank feature based on feature importance. The RFE method can be embedded in models and it removes the least important feature through recursions. To find the optimal number of feature that maximizes the performance, cross-validation is used to score different feature subsets and select the best scoring collection of features. At first, we input all the features to the models and set the number of features we want to keep. Then, the RFE will perform its feature elimination function to select a list of important features.

4.3 Fundamental Analysis

Fundamental analysis is a method of measuring intrinsic value of stock in the financial market(Majaski, 2019). It covers a wide range of external factors such as economic indicator, industry condition and

finance market condition. In this project, most of the fundamental analysis will focus on studying the impact of news sentiment since many previous researches fail to take news sentiment into consideration when analyzing external factors impacting bitcoin price.

4.3.1 News Sentiment Analysis

News per se are of less value if the sentiment cannot be properly quantified. To transform unstructured data to structured one, sentiment analysis is considered. The purpose of sentiment analysis is to determine the polarity of a piece of news by calculating a sentiment score based on the content. The sentiment score can be used for in-depth analysis, visualization and machine learning as well as deep learning algorithm. Theoretically, there are two basic natural language processing methods for sentiment analysis, one is machine learning based method and the other one is lexicon-based method. In this project lexicon-based method is selected because machine learning method requires massive amount of training data which is not readily available. Our team may need to create an ad hoc training data set in order to carry out the training and prediction. Lexicon-based method, by comparison, is more flexible and efficient since the only tool needed is a widely accepted lexicon. In this study, VADER (Valence Aware Dictionary and sEntiment Reasoner) dictionary is selected for several reasons. First, VADER is specifically attuned for social media text such as tweeter, reviews and news. Second, as a crowdsourcing lexicon based on Amazon's Mechanical Truck, the lexicon has access to the most recent social media content and self-reinforce its reliability. Third, the sentiment score produced by VADER lexicon is a number ranging from -1 to 1. If the sentiment is close to -1, it denotes a significantly negative sentiment and vice versa. VADER lexicon does not only reflect the subjectivity of news, but also shows the extend of polarity for each news. The sentiment score is vitally important in this project because of its interpretability regarding how social media, regulation and security issue impact bitcoin price.

Furthermore, news can be divided into different categories based on the context and purpose. For example, market, technical, regulation as well as privacy and security. Therefore, by studying different categories of news we can explore how bitcoin price will be impacted by regulation, policy and security. This will provide a more in-depth insight of news. Therefore, our group will conduct sentiment analysis on three categories of news: regulation, privacy and security, and the Wall Street Journal news (market and technical news). By studying each category of news, sentiment score is collected for each category on daily basis. If there is no relevant news on one day, the sentiment score on that day will be determined by following the sentiment score in the nearest day in the past. There will be three features extracted from news sentiment score, namely regulation sentiment, privacy and security sentiment as well as WSJ news sentiment.

4.4 Technical Analysis

(Majaski, 2019) points out that unlike fundamental analysis, technical analysis only considers volume and historical price as inputs. The assumption is that price has already reflected all other factors from fundamental analysis, so the impact of economic indicator, industry condition or finance market condition

can be omitted. The purpose of technical analysis is to capture the historical price movement pattern and make reliable prediction in the future.

4.5 Model Implementation and Hyper-parameters Tuning

4.5.1 linear regression

Multiple linear regression is the most common form of linear regression analysis. The multiple linear regression is used to explain the relationship between a one continuous dependent variable and two or more independent variables and response variable by fitting a linear equation to observed data. The independent variables can be continuous or categorical. Formally, the model for multiple linear regression, given n observations, is

$$y_i = \beta_0 1 + \beta_1 X_i 1 + ... + \beta_p X_i p + \varepsilon_i = X_i^T \beta + \varepsilon_i, i = 1, ..., n$$

The y_i means the price, and the parameters $\beta_0, \beta_1, ..., \beta_i$ is the features we have selected. The fitted values X_1, X_2,..., means the weight allocation of each feature. In this regression, calculating the different weightiness of feature to get the final price. So we can use the result to compare and analyze the historical price.

4.5.2 Support Vector Regression

Model Implementation: Support vector regression(SVR) is a version of support vector machine (SVM) for regression. It is widely used in data science industry since it can solve both linear and non-linear regression problems by identifying a hyper plane which maximizes the margin. For bitcoin price prediction, non-linear SVR would be more suitable. The non-linear model adds one of the kernel functions - linear, polynomial or Gaussian radial basis function in order to project non-linear data onto a higher dimensional feature space to make it possible to perform the linear separation.

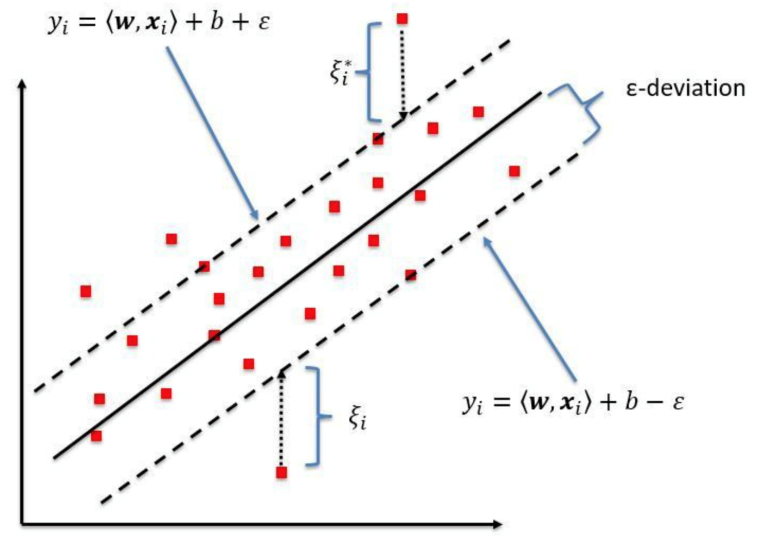


Figure 4. SVR Algorithm

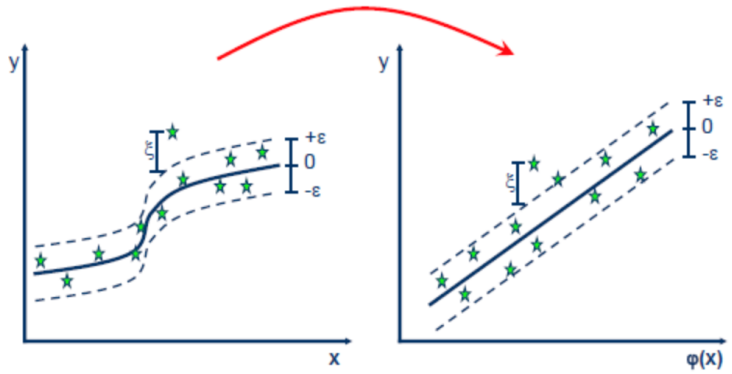


Figure 5. Kernel function to transform data into higher dimension and perform linear separation

Gaussian radial basis function (rbf) handles fluctuating data better than the other two kernels. Apply grid search method to tune hyperparameters of the rbf SVR model. E.g. select the combination of values of kernel coefficient (gamma), penalty parameter C of the error term and epsilon-tube (epsilon),which give highest prediction accuracy.

Hyperparameters Tuning: Grid search is a method to perform hyperparameter tuning. It is widely adopted in numerous algorithms to find the optimal set of hyperparameters that maximizes the performance. The idea of grid search is to try out all the possible combination of hyperparameters in a given range and select the best combination. Grid search is a good fit for SVR because of the complex nature of hyperparameters in SVR such as the kernel function, the penalty factor C and the value of gamma. Grid search also supports cross validation to enhance the output reliability. The number of cross validations is defined as "v" in the grid search function. Grid search with more rounds of cross validation tends to spend more time on producing the result with better performance. By utilizing Grid search in SVR, the optimal set of hyperparameters can be obtained at relative ease.

4.5.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a subclass of Recurrent Neural Networks (RNN), and it is designed to solve the short-term memory problem of RNN. It has internal mechanisms called gates, which regulates the flow of information. Purpose of implementing these gates is helping the model to learn which data in a sequence is important, and then decide to keep or drop that data. By doing so, it can pass significant information down the long sequences to make predictions. Figure 6 shows how a LSTM cell operates to keep or forget information. **Sigmoid** is an activation function, it squishes values between 0 and 1. Values closer to 0 means the corresponding information is not importance and it should be dropped while values closer to 1 indicates to keep the information. **Forget Gate** is to decide to drop or keep the information base on the results returned by sigmoid function. Previous hidden state and current input are then passed into a **Input Gate**, which contains sigmoid function to label the importance and a tanh function to squish values between -1 and 1 in order to regulate the network, none zero values after multiplication of sigmoid output and tanh output are transferred to next state. Next cell operation is to update the **Cell State** by

firstly multiply the input from previous cell with the output from the first sigmoid function and do a pointwise addition with the input gate. At last, the **Output Gate** pass the previous hidden state and the current input into a sigmoid function and then multiply with the newly modified cell state proceed by tanh function to output a hidden state for the use as an input of next LSTM cell. According to the mechanisms of LSTM, it should be a proper model for bitcoin price prediction since bitcoin price and relevant features are in time series (Nguyen, 2018).

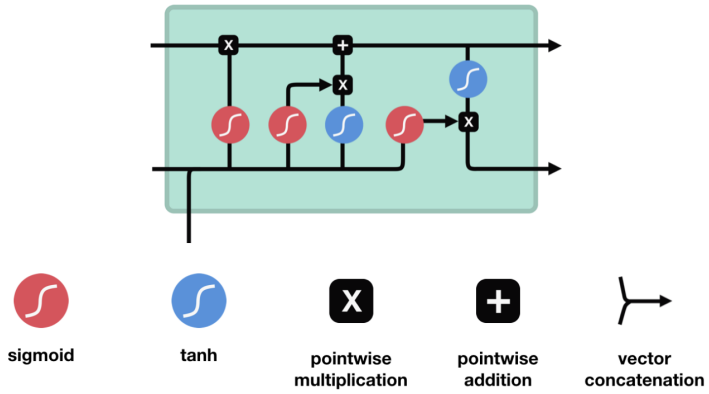


Figure 6. LSTM Cell and It's Operations

Model Implementation: A LSTM model is implemented for bitcoin daily close price prediction in this project. Model inputs are selected features with n-day time-stamp and outputs are the predicted bitcoin price of the next day. For example, to predict bitcoin price on day t by 3-day time-stamp, the inputs should be the feature values on day t-3, t-2 and t-1. Figure 7 shows working mechanism of this LSTM model.

Initial inputs of the LSTM model are selected according to the outcomes of feature importance analysis. They are: Open, close, high, low prices of bitcoin, volume, moving average, oil price, USD/EUR exchange rate and number of relevant tweets. Feature data are collected from 1 July 2015 to 16 March 2019, which contains 1355 days in total. Training set contains 1220 samples from 1 July 2015 to 1 November 2018, and testing set contains 135 samples form 2 November 2018 to 16 March 2019. Validation method is not implemented due to unstable performance produced by small amount training sample.

Hyperparameters Tuning:

Further feature selection and hyperparameters tuning are necessarily to be completed by trying different combination of feature inputs, as well as tuning the learning rate, number of hidden layer and hidden node, choosing most appropriate optimizer and loss function and run the model for a reasonable amount of epochs. Moreover find out a time-stamp that is able to optimize the model performance is significantly important.

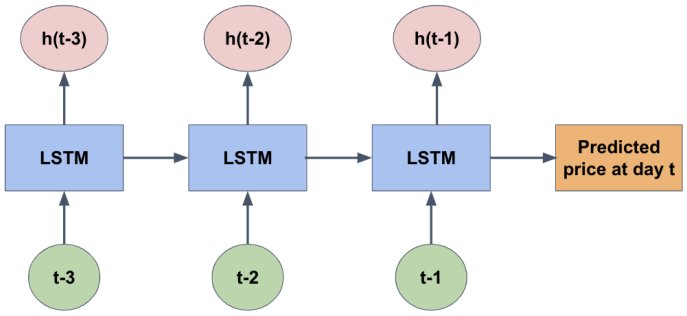


Figure 7. LSTM Model Structure implemented in this project

5 RESOURCES

5.1 Hardware

The software tools will run on own computers, the detail information has shown below: Macbook Pro(13-inch,2017,Two Thunderbolt 3 ports) Processor: 2.3GHz Intel Core i5 Memory: 8 GB 2133 MHz LPDDR3 Graphics: Intel Iris Plus Graphics 640 1536 MB

5.2 Softeware

The following software tools have used in our project: Jupyter notebook,Pycharm,TensorFlow,Tableau,Microsoft Word,Microsoft Excel

5.3 Other Resources

Here are some other resources we have used in data collection and data analysis: Twitter news,Bitcoin Magzine,stock market platfrom and digital currency platform.

5.4 Roles and Responsibilities

6 MILESTONE/SCHEDULE

Kick off and design work plan: The project started on 06/03/2019. First, we should develop the plan for the entire project, including the duration of the entire plan, the starting data for each phase, and so on.

Define our aim and scope: At this stage, we conducted background research and methodological research. And confirm our goal and final deliverable.

Literature Review summary: collect papers and complete literature review. By the end of this phase, the main deliverable are literature reviews

Proposal Report Due: This is a milestone for the project, we will present our Methodologies and time-line.

Define hypothesis for each other: find the relationship between factors and bitcoin price.

Perform statistical analysis and visualization Apply machine learning algorithm: Compare different models. The main deliverable is an algorithm for finding the relationships between them.

Name	Roles	Responsibility
Chenxiao Cai	Leader, trouble shooter	1. Managing progress and direction of the project, weekly planning 2. Serve as a bridge between the group and tutor, responsible for communicating and discussing progress and questions with tutor on weekly basis. 3. Data collection including gold price, oil price and exchange rate. 4. Perform statisitcal analysis and feature importance analysis to determine the relative importance of each factor in the candidate data list. 4. Develop python news scrapper to mine news from the Wall Street Journal(WSJ). 5. Conduct news sentiment analysis on regulation news, privacy and security news as well as WSJ news, and data mining to generate intuitive insights and visulization from the news. 6. Design and develop linear regression and SVR model to predict bitcoin price by choosing different sets of features. 7. Report writing and auditing.
Chuhan Wang	Trouble shooter	1. Complete the tasks assigned by the group. 2. Complete data collection - (all bitcoin related features) and preparation for modelling. 3. Perform statistical analysis - correlation coefficient and relative importance analysis. 4. Implement LSTM model for bitcoin price prediction. 5. Read relevant news and analyse how different news affect bitcoin price. 6. Report writing and auditing.
Xingchen Zhou	Trouble shooter	1. Complete the task assigned by the group. 2. Prepare the visualization model. 3. Complete data collection, prepare for data analyze and modelling. 4. Collect and complete related literature collection. 5. Document literature review. 6. Report writing and auditing.
Sihong Huang	Trouble shooter,Mom-tracker	1.Complete the task that assigned by the team. 2.Recording the content about meeting of minutes. 3.Monitor the progress of the current stage, provideing the Idea and opinions for the task. 4.Data collection including bitcoin news,mining difficulty. 5.Report writting and auditing.

Figure 8. Roles and Responsibilities

Progress Report: This is a milestone for the project. This report will describe the progress of the project and the steps that will be taken during the following phases.

Test the model by comparing: We will test model for many times and we will gradually optimize the performance of the model to improve its performance.

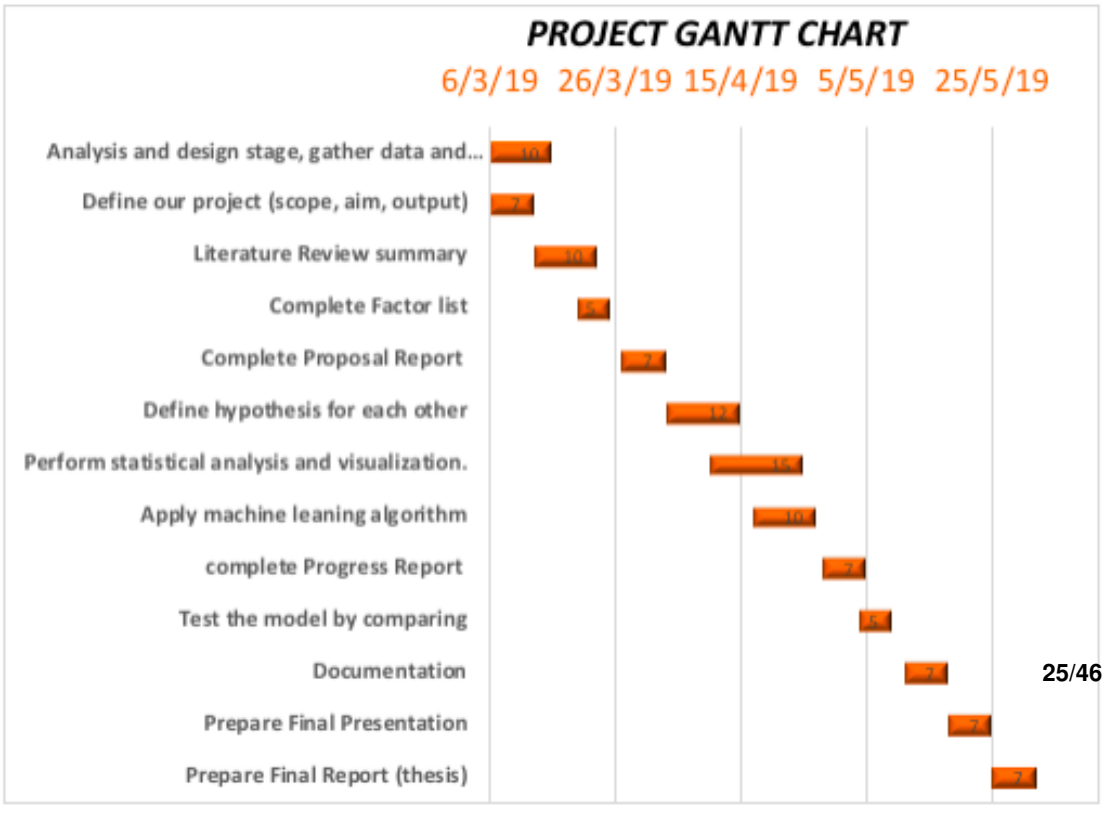
Documentation: At this stage, some instructions will be provided. The instructions include some readable functions, grades, etc.

Final report: The final report and report are milestones of the project. The presentation will discuss the content of the project and results.For the final report, it will explain more details about the methods of comparison.

The detailed Gantt chart is provided below:

Milestone	Tasks	Start Date	Duration	End Date
Week-1	Kick-off Meeting Analysis and design stage	06/03/2019	10	16/3/19
Week-2	Define our project (scope, aim, output)	06/03/2019	7	13/3/19
Week-3	Literature Review summary	13/03/2019	10	23/3/19
Week-4	Complete Factor list	20/03/2019	5	25/3/19
Week-5	Proposal Report Due	27/03/2019	7	3/4/19
Week-6	Define hypothesis for each other	03/04/2019	12	15/4/19
Week-7	Perform statistical analysis and visualization.	10/04/2019	15	25/4/19
Week-8	Apply machine leaning algorithm	17/04/2019	10	27/4/19
Week-9	Progress Report Due	28/04/2019	7	5/5/19
Week-10	Test the model by comparing	04/05/2019	5	9/5/19
Week-11	Documentation	11/05/2019	7	18/5/19
Week-12	Final Presentation	18/05/2019	7	25/5/19
Week-13	Final Report (thesis)	25/05/2019	7	1/6/19

Figure 9. Milestone



7 RESULTS

7.1 Feature Importance

The result of feature importance and feature selection is provided in Figure 11:

	SVR with Recursive Feature Elimination (Top 5)	Random Forest with forward selection(Top 5)	Random Forest	XGBoost	Mutual Information	Linear Regression	Multi Various Regression	Ranking Avg
3 Days Moving Average	Selected	Selected	1	1	1	1	1	1
Litecoin Price	Selected	Selected	5	2	5	4	2	3.6
9 Days Moving Average	Selected	FALSE	2	8	3	2	6	4.2
14 Days Moving Average	Selected	FALSE	3	7	4	3	5	4.4
Minning Difficulty	Selected	Selected	4	12	2	10	4	6.4
USD/EUO Exchange	FALSE	FALSE	7	5	9	5	11	7.4
Volunm	FALSE	FALSE	9	3	12	11	7	8.4
Number of Tweets	FALSE	Selected	14	4	14	6	9	9.4
Viki View	FALSE	Selected	11	10	15	9	3	9.6
Address	FALSE	FALSE	15	6	13	8	8	10
Gold Price	FALSE	FALSE	6	9	10	12	13	10
Privacy and Security Sentiment	FALSE	FALSE	10	11	7	16	10	10.8
Regulation News Sentiment	FALSE	FALSE	13	14	6	13	12	11.6
Oil Price	FALSE	FALSE	12	16	11	7	14	12
Transcation	FALSE	FALSE	8	13	16	14	16	13.4
WSJ News Sentiment	FALSE	FALSE	16	15	8	15	15	13.8

Figure 11. Result Summary of Feature Importance

It comes to our attention that it is 3 days moving average that stands out in terms of feature importance ranking in all measurements including random forest feature importance, XGBoost feature importance, mutual information regression, linear regression as well as multi various regression. When it comes to recursive feature elimination and forward selection, 3 days moving average is also selected in both feature selection procedures, demonstrating that 3 days moving average is the most important factors influencing the bitcoin price movement among all candidate factors.

Apart from 3 days moving average, Litecoin price, 9 days moving average, 14 days moving average, mining difficulty as well as USD/EUO exchange rate also reflect a strong statistical relationship with bitcoin price. Other factors, by comparison, is relatively less statistically important.

Variables	Raw RelWeight Rescaled	RelWeight	Rank by Raw Weight
Volume	0.0476	4.8560	8
Num_bitcoin_address	0.1062	10.8441	6
mining_difficulty	0.1036	10.5742	7
Weighted_litecoin_price	0.2477	25.2927	1
Oil_price	0.1145	11.6905	4
usdEuro	0.1345	13.7369	2
Num_Wiki_view	0.1069	10.9158	5
Num_tweets	0.1184	12.0898	3
R^2	0.9794		

Figure 12. Result Summary of Relative Importance Analysis

Due to limited number of factors that can be input to the system to calculate relative importance weights, we selected some fundamental factors beside technique factors. The result (12) shows litecoin price contributes most weight to bitcoin price - 0.2477, which is significantly higher than other factors. USD/EUR ratio and number of relevant tweets takes the 2nd and 3rd place - 0.1345, 0.1184 respectively. By 95 % confidence interval and significant test, all of the factors are correlated to bitcoin price and we are satisfy with this outcome as the R square values is 0.9794, which means these factors are able to explain bitcoin

price exchange to some extent.

7.2 Linear Regression and SVR Model

For prediction, three separate experiments are delivered to find the best set of features that maximize the model performance. There are 1,335 instances in total. The 1,335 instances are divided into 1,210 training data (from 1st June 2015 to 30th November 2018) and 125 testing data(from 1st December 2018 to 16th March 2019). In addition, three metrics are selected to evaluate the performance of the machine learning model. They are Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). There are two reasons why MAPE, MAE and RMSE are chosen for this problem. First, these measurements are widely used in previous research paper, so in the matters of comparability concerned, we ought to use the same performance metrics to determine whether our prediction outperforms others or not. Second, the bitcoin prediction is a regression problem rather than classification problem, empirically speaking MAPE, MAE and RMSE are fit for regression problem because all three metrics measures to what extend does the predication matches the actual. A low number in MAPE, MAE and RMSE denotes performance excellence while a high MAPE, MAE or RMSE flags out that the model may need further testing and optimization. The performance summary is provided below:

		Feature	MAPE	MAE	RMSE
Linear Regression	Experiment 1	3 days moving average, 9 days moving average	2.92%	119.27	185.27
	Experiment 2	3 days moving average, Litecoin price, 9 days moving average and 14 days moving average.	2.87%	118	186
	Experiment 3	3 days moving average	2.85%	127.29	199.74
SVR	Experiment 1	3 days moving average, 9 days moving average	3.67%	119	183.89
	Experiment 2	3 days moving average, Litecoin price, 9 days moving average and 14 days moving average.	0.94%	195	264.18
	Experiment 3	3 days moving average	2.27%	129	201.5

Figure 13. Result Summary of Linear Regression and SVR

Three experiments are conducted for each machine learning algorithm. Based on the performance metrics, the performance of experiment 1 and experiment 2 are quite close, so it is difficult to determine which feature combination produces best result for linear regression. When it comes to SVR, experiment 1 outperforms experiment 2 and 3 in terms of all the valuation metrics. Thus in this dataset, selecting 3 days moving average and 9 days moving average gives the best result for bitcoin price prediction under SVR.

The prediction vs actual line charts for each experiment under linear regression and SVR are provided below:

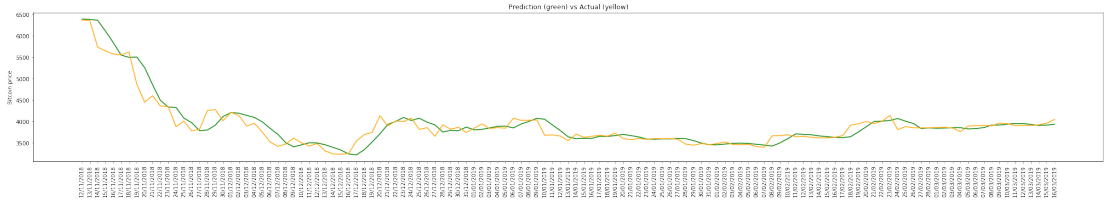


Figure 14. Linear regression experiment 1 prediction vs actual

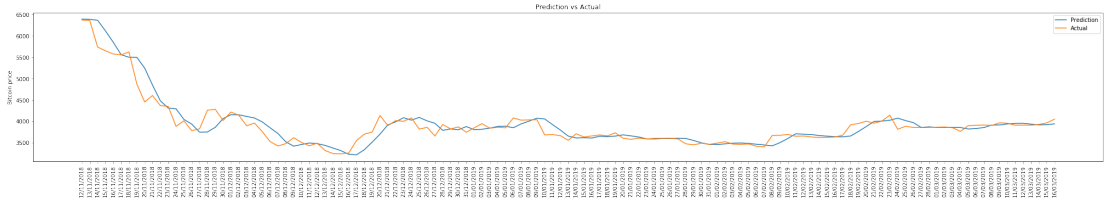


Figure 15. Linear regression experiment 2 prediction vs actual

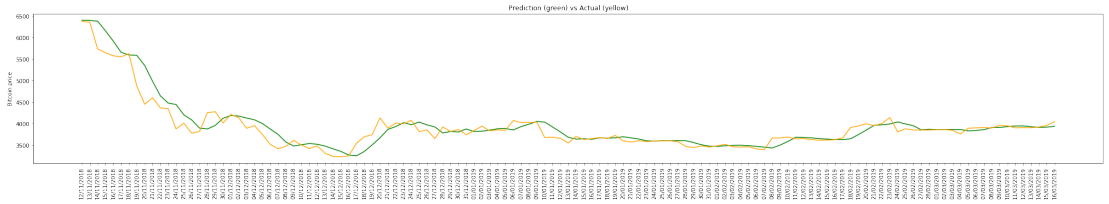


Figure 16. Linear regression experiment 3 prediction vs actual

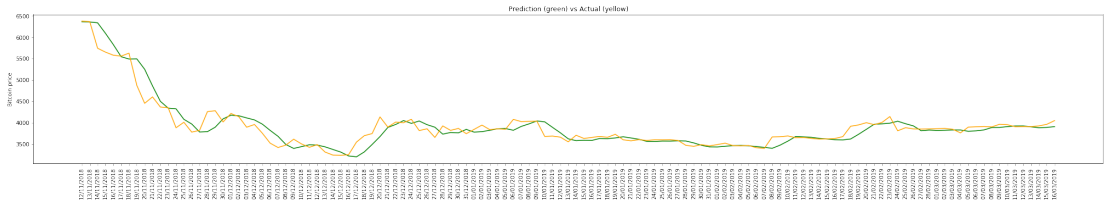


Figure 17. SVR experiment 1 prediction vs actual

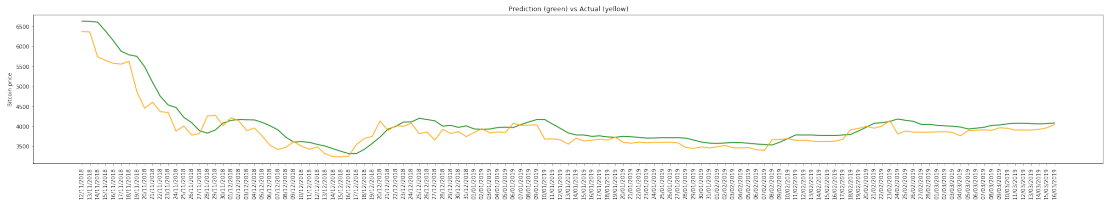


Figure 18. SVR experiment 2 prediction vs actual

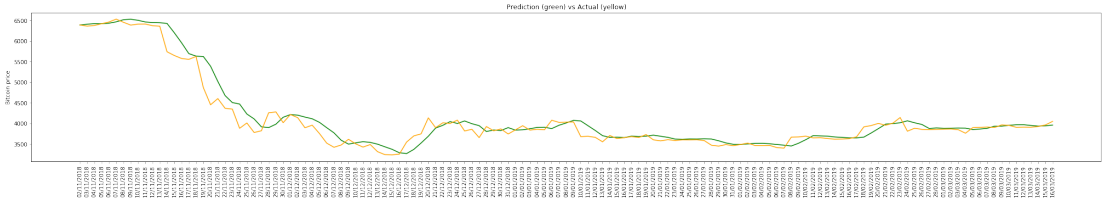


Figure 19. SVR experiment 3 prediction vs actual

7.3 LSTM Model

Figure 20 shows prediction results of test dataset by implementing different combinations of input features by 10-day time stamp. After finding out the best input features, Figure21 displays model performance of different time stamps while remaining input features unchanged. Open, close, high, low prices, volume changes in percentage and moving average is the best inputs features combination and 10-day is tested as the most proper choice of time stamp. Figure 22 is the plotting of actual and predicted bitcoin price of 125 days.

Experiments		Results	Open	Close	High	Low	Volume(%)	9-day Moving Average	USD/EUR	Number Tweets	Oil Price
Model 1	Test RMSE	156.12	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Test set classification	56.45%									
Model 2	Test RMSE	178.46	✓	✓	✓	✓	✓	✓	✓	✓	
	Test Classification Acc	56.06%									
Model 3	Test RMSE	179.43	✓	✓	✓	✓	✓	✓	✓		
	Test Classification Acc	62.10%									
Model 4	Test RMSE	115.17	✓	✓	✓	✓	✓	✓			
	Test Classification Acc	62.10%									
Model 5	Test RMSE	202.1	✓	✓	✓	✓	✓				
	Test Classification Acc	54.03%									

Figure 20. Prediction results of different input feature combinations

Time-stamp		Results	Open	Close	High	Low	Volume(%)	n-day Moving Average
3-day	Test RMSE	174.77	✓	✓	✓	✓	✓	✓(3-day)
	Test set classification	48.70%						
10-day	Test RMSE	115.17	✓	✓	✓	✓	✓	✓(9-day)
	Test Classification Acc	62.10%						
14-day	Test RMSE	167.44	✓	✓	✓	✓	✓	✓(14-day)
	Test Classification Acc	50.83%						

Figure 21. Prediction results of different time stamps

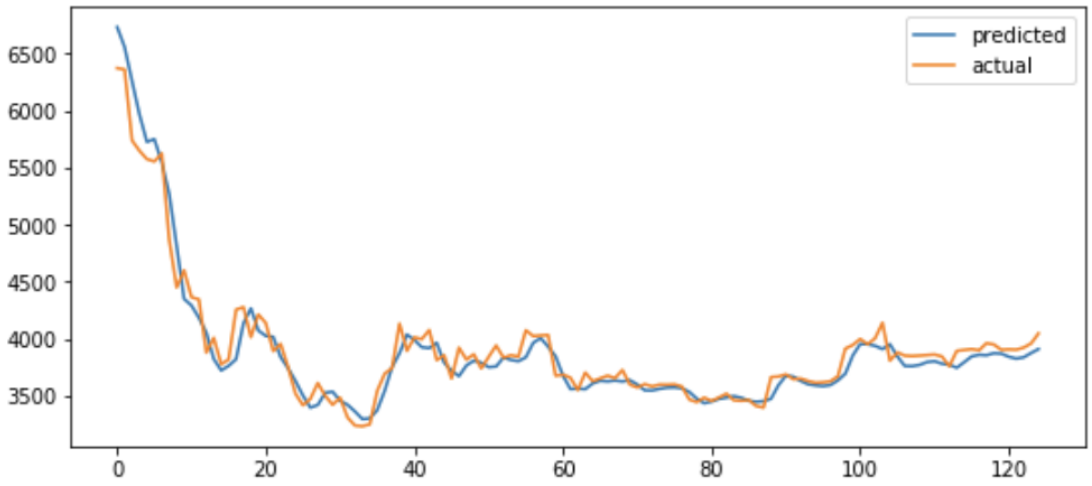


Figure 22. Acutal and Predicted bitcoin close price

7.4 News

7.4.1 Regulation and Policy news

Figure 23. Regulation news word cloud

The word cloud provides visualization of the most frequently appeared word in regulation news. Bitcoin, SEC, regulation, blockchain, new and government are the words with most appearance. It can be concluded that most news is related to SEC, an US government body attempts to regulate the blockchain in order to protect investor benefits. The result of word cloud is aligned with the regulation news regional distribution, that is, U.S owns most regulation news headlines.

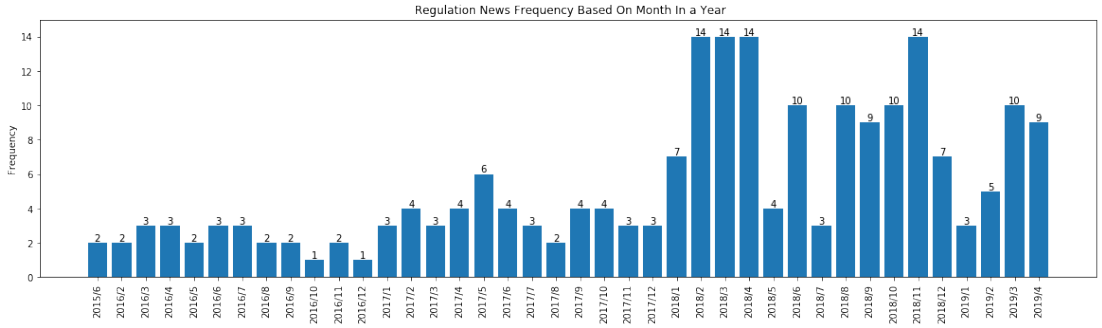


Figure 24. Regulation news frequency

Figure 24 provides information about monthly regulation news frequency from June 2015 to April 2019. Apparently, it can be inferred that the number of regulation news peaked at the first quarter of 2018 and the second half of 2018. There is also several regulation news about bitcoin in the past 2 months.

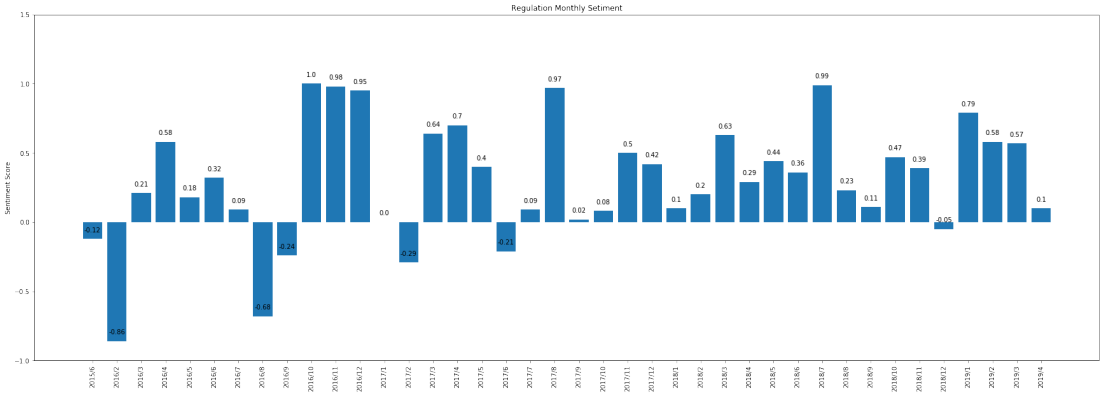


Figure 25. Regulation news monthly sentiment

The monthly regulation sentiment shows that bitcoin regulation introduced by governments in the globe is generally positive. This positive pattern is particularly strong in the last quarter of 2016 (October to December). The recent regulation sentiment also shows a positive pattern.

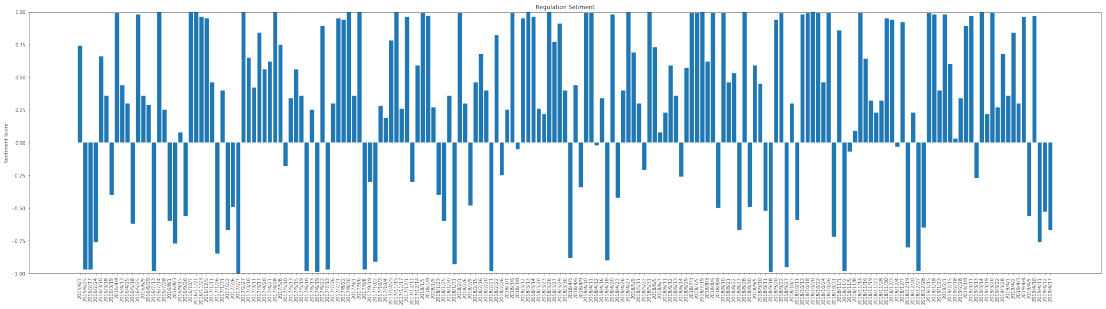


Figure 26. Regulation news daily sentiment

The regulation news daily sentiment is provided below in Figure 26. The overall trend reveals that the

regulation news tends to be positive in the past with few exceptions. It shows that governments all over the world plans to regulate the use of bitcoin so that investor benefits can be guarded, and illegal crossings can be eradicated.

The bar chart (Figure 27) shows the regional distribution of regulation news. The U.S. takes the lead at 117, which is more than EU, China, Japan and South Korea combined.

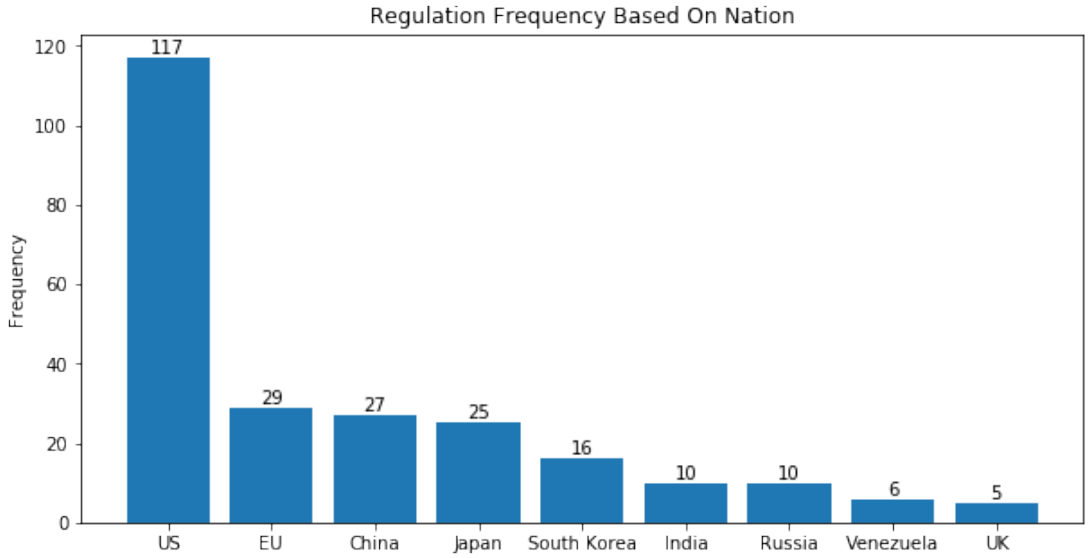


Figure 27. Regulation news country distribution

The line chart together with the bar chart (Figure28 and Figure 29) gives information about the count of positive and negative regulation news over time. In general, there are more positive regulation news than negative ones. There was a surging number of positive regulation news between February 2018 and March 2018 followed by a fluctuation. In addition, the number of negative regulation news is also partially associated with the number of positive news, given that between February 2018 and Dec 2018 the number of negative regulation news was in good part determined by number of positive regulation news.

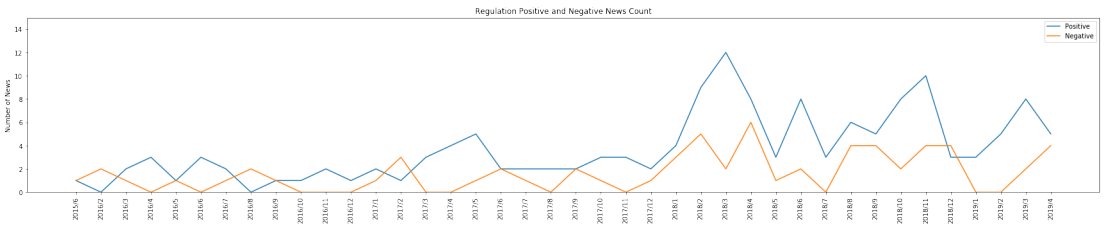


Figure 28. Regulation news positive and negative count

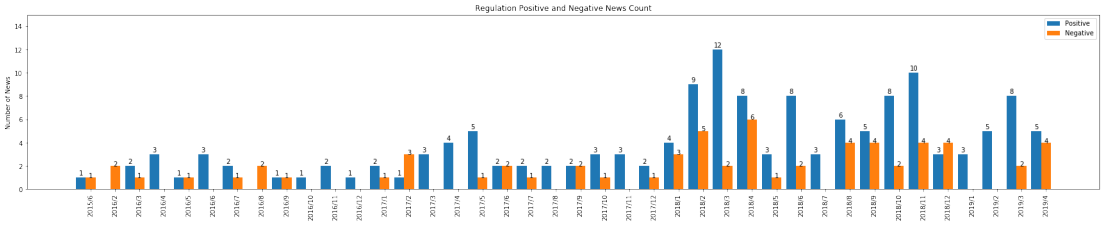


Figure 29. Regulation news positive and negative count bar

The graph shows the bitcoin price movement in relation to regulation sentiment. Generally, there is no clear evidence to identify a plausible statistical relationship between bitcoin price and regulation. So overall no consistent patterns can be found to link bitcoin price with regulation sentiment.

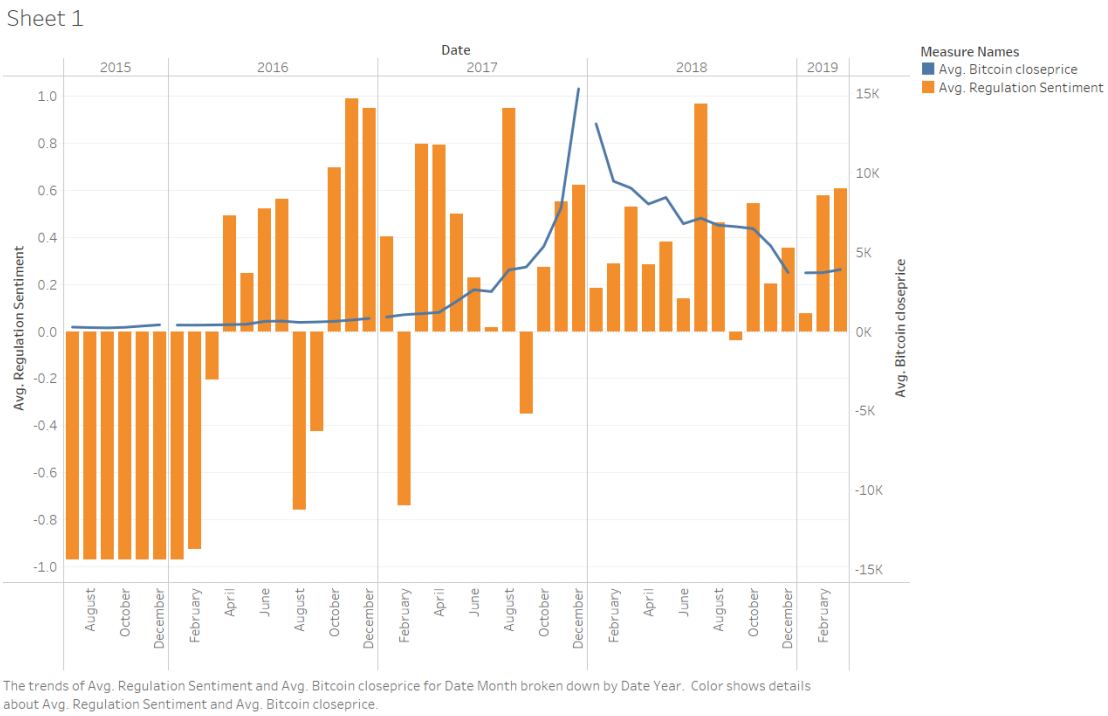


Figure 30. Visualization of regulation sentiment and bitcoin price

7.4.2 WSJ news

The word cloud for WSJ news shows a slightly different insight. Since the WSJ specializes in market news, the most frequently appeared words are exchange, trading, cryptocurrency and market. Moreover, China and the U.S are the top countries associated with bitcoin market in the globe.

Figure 31. WSJ news word cloud

The WSJ news monthly sentiment shows that the market and technical news sentiment tend to be neutral in the past. Yet some outliers can not be omitted, for example, the sentiment in August 2015 was highly negative while the counterparts in October and November 2015 reached record high. Noticeably, the NSW news sentiment was considerably low in February this year.

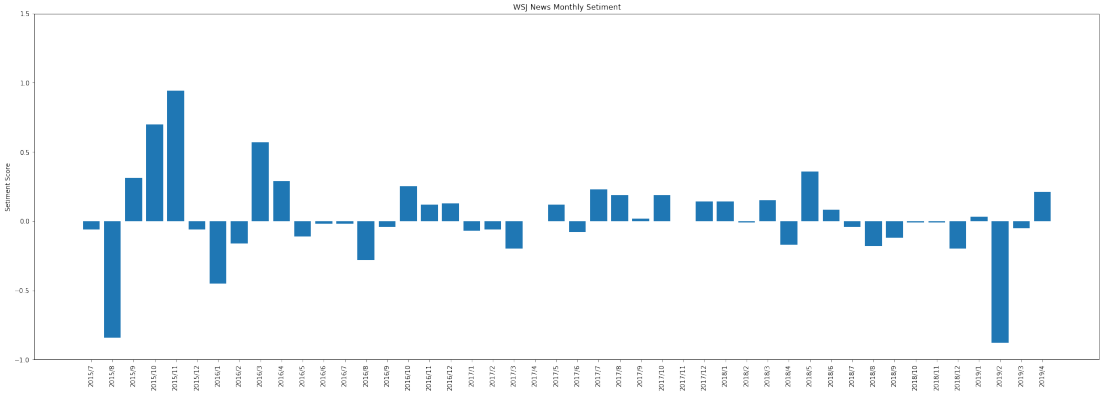


Figure 32. WSJ monthly sentiment

The line chart and two bar charts below jointly give a comprehensive view of number of positive and negative WSJ news over time. The most striking pattern is that the number of positive news embraced a surge between November 2017 and December 2017. Similarly, the number of negative news also saw a significant rise accordingly.

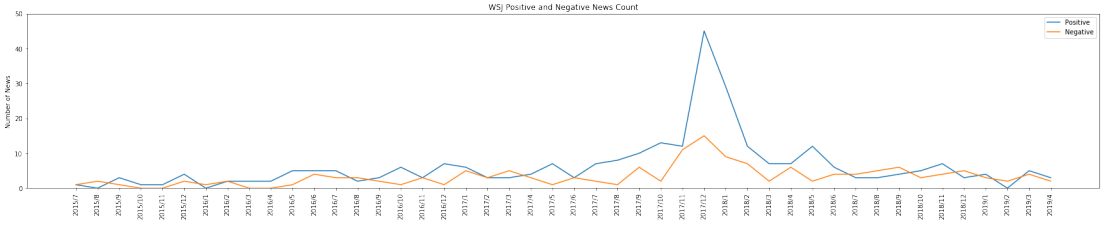


Figure 33. WSJ positive and negative news count

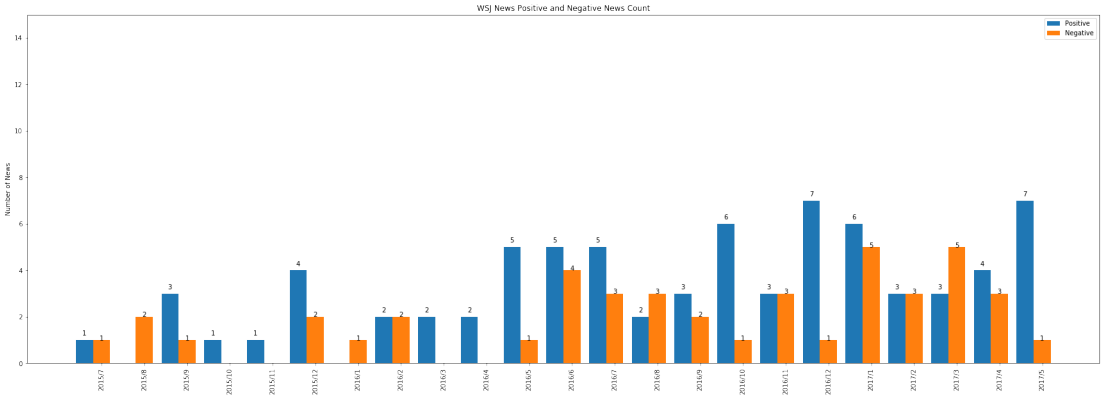


Figure 34. WSJ positive and negative news count (July 2015 to May 2017)

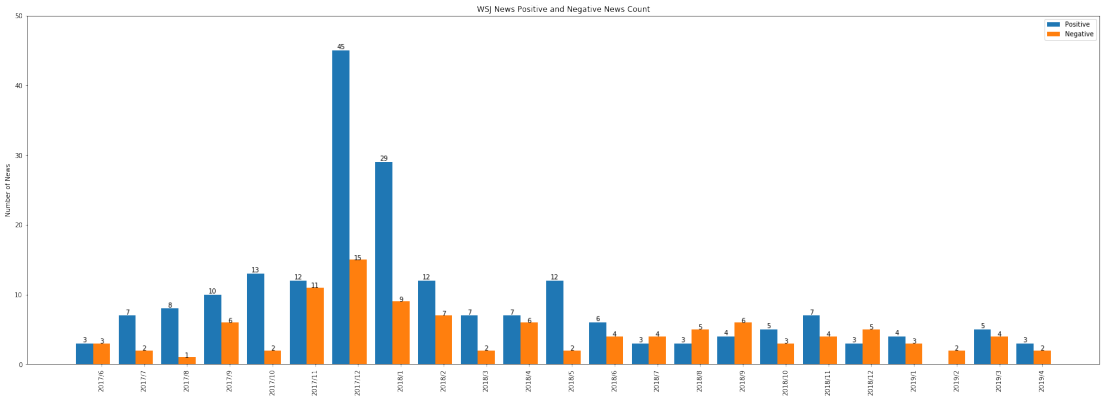


Figure 35. WSJ positive and negative news count (June 2017 to April 2019)

Most WSJ news are concerned with the market and technical aspect of bitcoin. It turns out that the sentiment tends to be natural when the bitcoin price witnessed a surge followed by sharp avalanche between October 2017 and April 2018. Moreover, the WSJ shows an ever-pessimistic view on bitcoin recently (from December 2018 to March 2019).

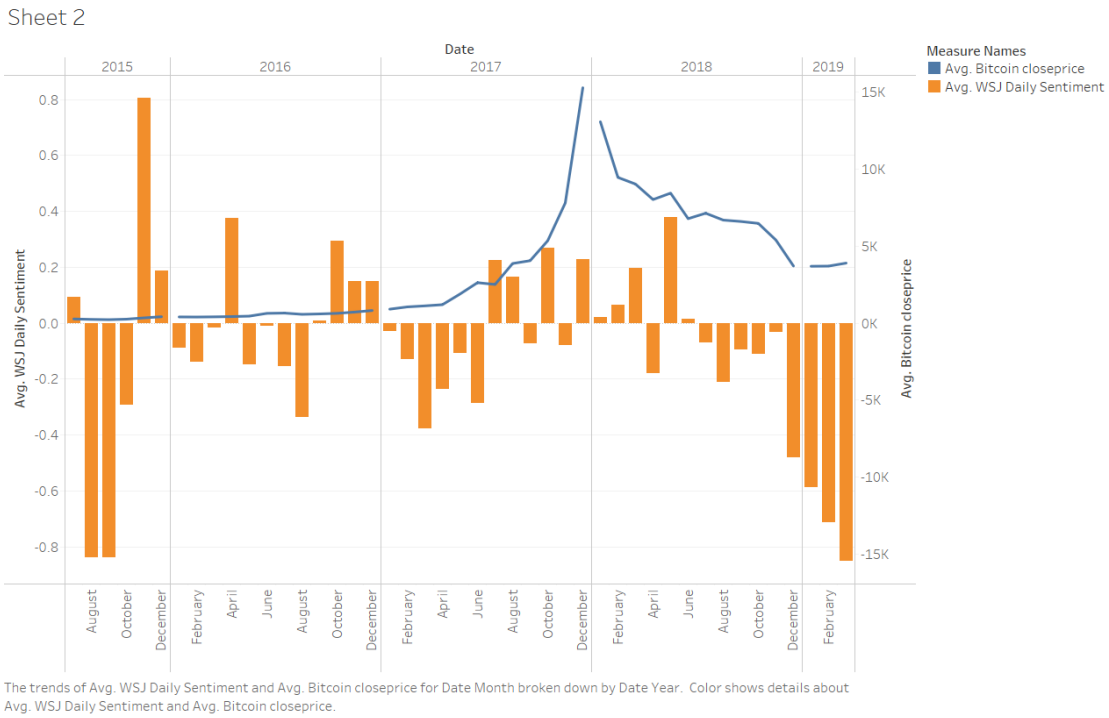


Figure 36. Visualization of regulation sentiment and bitcoin price

7.4.3 Privacy and Security news

Figure 37. Privacy and security word cloud

When it comes to word cloud for privacy and security news, it comes to our attentions that privacy and security news concerns more about user transaction and exchange as well as blockchain security. The privacy and security news daily sentiment are provided below:

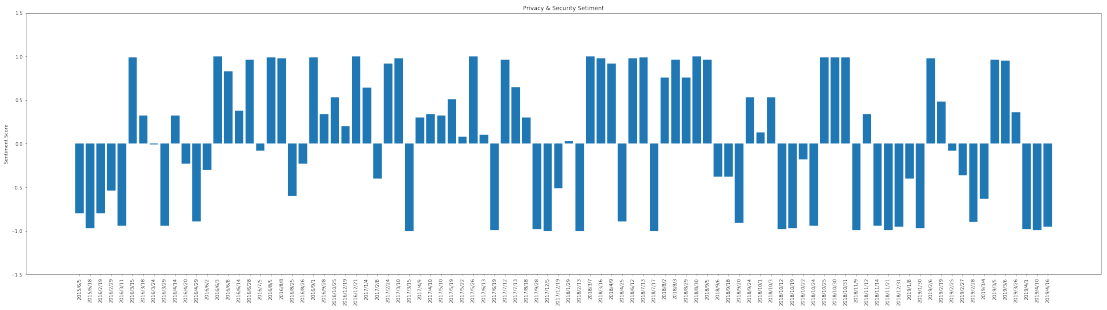


Figure 38. Privacy and security news monthly sentiment

The historical pattern shows that the privacy and security sentiment has been fluctuated in the past four years. Between June 2015 and March 2016, the sentiment was consistently negative, in a sharp comparison to that in April 2017, May 2017, August 2018 and September 2019. Noticeably, the privacy and security sentiment are negative in the past 3 successive months.

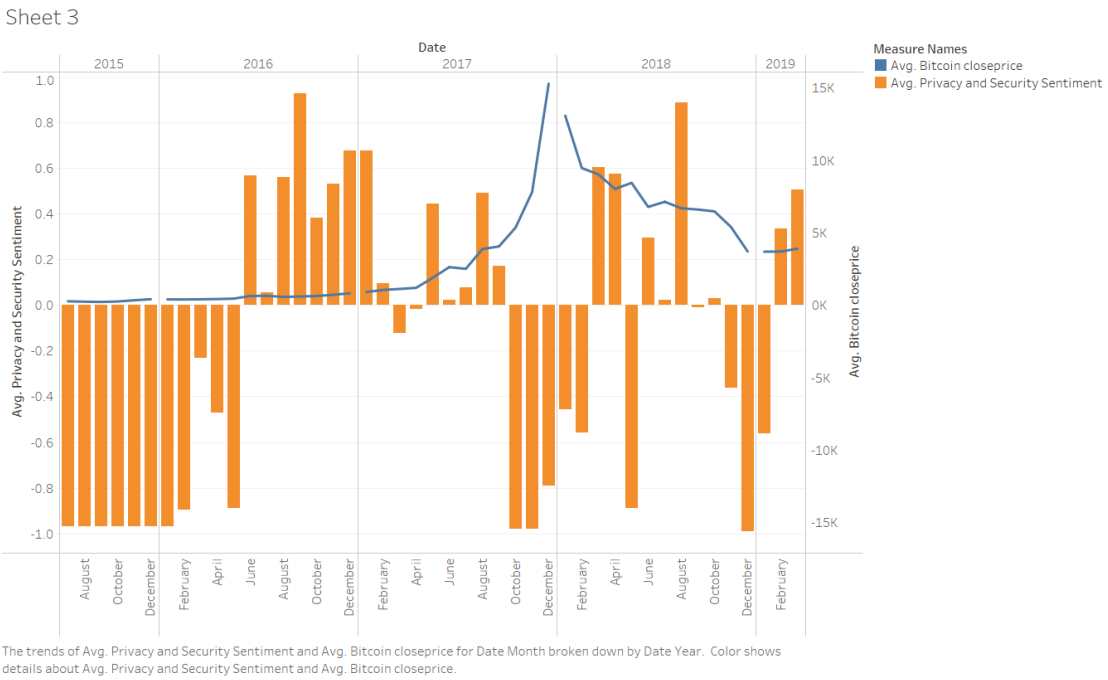


Figure 39. Visualization of privacy and security sentiment and bitcoin price

The relation between bitcoin price and privacy and security news sentiment is akin to that of regulation sentiment. However, at certain observation point, there is a clear linkage. Starting from October 2017 to December 2017, the leap in bitcoin price was paired with highly negative privacy and security sentiment. In addition, at the early stage (2015 to 2016), the public perception about bitcoin security and privacy remained negative for 11 successive months until June 2016.

8 DISCUSSION

8.0.1 External factor importance

The short-term, mid-term and long-term bitcoin historical price together with Litecoin price are the top 4 features in terms of feature importance. The finding is aligned with the assumption of technical analysis, that is, the impacts of external factors have already embedded in the price itself. No further data is required to capture the bitcoin movement pattern. 3 days moving average is the most important feature because it shows the average closing price of bitcoin in the past 3 days, which is significantly relevant to the price movement in the next day. Although Litecoin is another type of cryptocurrency, its characteristics are highly akin to that of bitcoin. Thus, Litecoin and Litecoin are highly interdependent. If investors lose confidence in Litecoin, it is highly likely that bitcoin price will suffer accordingly.

Mining difficulty is also a vitally important external factors since it subtly measures the supply of bitcoin. Supply and demand are the fundamentals of every single goods and services in economics and bitcoin is no exception. Empirically speaking, bitcoin price will increase as mining difficulty rises. Moreover, volume is the 7th most important factors that measures bitcoin supply as well. Active addresses, number of relative tweets and Wikipedia views could be all considered relevant to demand of bitcoin and they are highly correlated to bitcoin price. These factors are able to reflect the confidence of bitcoin from the investors/the public as well as bitcoin popularity. Therefore, it can be concluded that bitcoin supply and demand is one of the determining factors that influence bitcoin exchange.

Moreover, exchange rate is another key external factor. In this project, USD/EUO exchange rate is selected because both Euro and US dollars contribute to a large portion of bitcoin transactions in the globe. The USD/EUO exchange rate reflects the demand of bitcoin. In the case of Euro or US dollar depreciation, investors holding these currencies are prone to adjust their investment portfolio by converting Euro and US dollar to other investments, so bitcoin is one of their viable options.

The news sentiment is not as important as the other external factors discussed above. The rank of news sentiment is directly attributed to three reasons. Frist, the number of bitcoin news is solely depending on the topic popularity in the public and market. If no big issue occurs, press is reluctant to release news about bitcoin. Therefore, bitcoin news is not readily available on daily basis. Second, the assumption of sentiment analysis adopted in the project. Since bitcoin news is not available every day, the average daily news sentiment for the day without any bitcoin news will follow the bitcoin news sentiment in the nearest day from the past. Third, lexicon-based sentiment analysis is not perfect, meaning that some of the sentiment scores do not provide a holistic assessment of the news due to the complex subtle meaning and context in news. Noticeably, regulation news sentiment and privacy and security news sentiment are slightly more important than WSJ news sentiment. This may infer that investors' concerns more about legal and security factors than market and technical factors. Further discussion and analysis will be engaged in greater detail in the following sections.

8.0.2 Regulation news

a) China bitcoin ban The bank of China officially banned bitcoin trading on 4th September 2017. The legal news collected in September 2017 are all related to China bitcoin ban. In total there are three regulation news in September 2017 with sentiment score 0.9978, -0.9712 and -0.296. The first one with 0.9978 is concerned with expressing an opinion of China's over-regulation on blockchain and ICOs. In the end the author argues that there is no need to worry about blockchain development in China for several arbitrary reasons. The other two with negative sentiment, by comparison, focus on stating the fact that China banned bitcoin trading in the first week of September 2017 and discuss the implication as well as potential panic triggered from this ban. Overall, from legal perspective, the sentiment in September 2019 is moderately negative. As a result, on 4th September 2017, the day when bitcoin was banned in China, bitcoin price fell from 4603.60 US dollars to 4277 US dollars followed by a further price decline in the following week. On 14th September 2017, 10 days after the China bitcoin ban, the price dropped by 30 percent to 3227 US dollars.The impact lasted for about two weeks until 17th September 2017.

Figure 40. Impact of bitcoin ban by China (Trendview,2019)

b) Twitter ban bitcoin ads On 27th March 2018, twitter abruptly banned all cryptocurrency ads. As a result, bitcoin price fell from 8152.26 US dollars to 7742.11 US dollars. The ad bans further deteriorated bitcoin price for the next 5 successive days with a 19.65 percent drop to 6813.52 US dollars on 1st April 2018.

Figure 41. Impact of ad ban by twitter (Trendview,2019)

c) Regulation and bitcoin crisis at the beginning of 2018 Although there is no systematic pattern between regulation news sentiment and bitcoin price. At the time of bitcoin crisis at the beginning of 2018, there

was a negative relationship between the two. From February 2018 to April 2018 there are 14 regulation news on each month with overall positive sentiment. Intuitively, three regulation news are released one month right after the bitcoin price avalanche in January 2018. To dig it deeper, the regulation discussed in the news during the three-month window is related to introduce new policy and regulation to regulate bitcoin transaction and mitigate bitcoin investment risk. That is why there were more far more positive legislation news than negative ones between February 2018 and April 2018. The origin of the regulation ranges from the U.S, EU, China, Hong Kong, Japan, South Korea to India. This is an unprecedented phenomenon because numerous countries enforced massive regulation and policy on bitcoin one month after the bitcoin crisis. Thus, it can be inferred that those regulations are primarily aimed at tacking the problems stemming from the crisis in January. Therefore, it concludes that regulation nowadays tends to be corrective other than preventive actions against bitcoin. Bitcoin price is the determining factor of bitcoin regulation and policy. In addition, there is a lagging response from the legislators since the government needs to spend time on understanding the problem, identifying factors contributing to the problem, choosing solutions from a range of proposals and passing the proposal. The lagging response explain why many regulations were put in practice one month after witnessing bitcoin crisis at the beginning of 2018. Therefore, legislation sentiment is not always an accurate indicator of bitcoin price, however, at the time of big event, legislation sentiment may provide meaningful insights on the bitcoin dynamics.

8.0.3 NSJ news

The most striking pattern in WSJ news sentiment is that the number of positive news embraced a sharp increase in December 2017 when bitcoin price reached a record high. However, the corresponding sentiment in December 2017 does not show a highly positive figure. The reason being that bitcoin price leaped to its record high on 16th December 2017 yet plunged sharply the day after. We found that from 1st December 2017 to 16th December, there are 41 bitcoin news published by WSJ, among which 12 are negative, 4 are neutral and 25 are positive. From 17th December 2017 and 31st December 2017, there are 19 bitcoin news. Only 5 of them are highly positive, the reaming 14 are either neutral or negative. Another apparent pattern is that between May 2018 and April 2019, the monthly sentiments are all negative while bitcoin price dropped sharply and eventually flatten out at 5,000 US dollars. Therefore, WSJ news sentiment can be used as an indicator for bitcoin price movement only if there is a price surge or drop. In such case, WSJ news sentiment can be valued asset for investors to make investment decision.

Noticeably, unlike that of regulation news, there is no lagging response for WSJ news. This is justified by the fact that the number of bitcoin news by WSJ surged on the exact same day when bitcoin price increased. Because WSJ focus on market news, so as long as there is big issue about bitcoin that worth publishing, the news will be released immediately.

8.0.4 Privacy and Security news

Overall the relationship between bitcoin price and privacy security sentiment is erratic in the long term. However, at the time of bitcoin price surge and crisis, some patterns can be observed. Between October 2017 and December 2017, the privacy security sentiment was highly negative while the bitcoin price increased dramatically. The privacy security news during the 3-months window talks about several privacy security issues. To begin with, the US government's determination to enforce new privacy security requirement that controls the traceability of each transactions so that bitcoin trader will never evade tax liability for bitcoin transaction. Second, one new revealed the fact that bitcoin was so popular that it became the favorite target of global DDoS attacks. The third news reports that a South Korean exchange platform called Youbit closed down due to a second hack in 2018. It turns out that once the popularity of bitcoin rises, it grabs the attention from both the government body and hackers. Similarly, on January 26th, 2018, one of the largest hacks in cryptocurrency history took place, as a result, over 530 million USD worth of NEM tokens were stolen by hackers. This unprecedented hacking leads to negative sentiment in both January and Feburay2018, which may also partially attribute to the sharp bitcoin price decline during the same time period. Therefore, privacy security sentiment is not always corelated to bitcoin price movement. Yet at the time of privacy security breach, it will significantly impact the bitcoin price since investors tend to lose confidence in bitcoin if security and privacy cannot be reserved.

8.0.5 News summary

In brief, there is no consistent correlation between news sentiment and bitcoin price over time. However, in the case of big event such as bitcoin price surge or crisis, news sentiment is one of the determining factors influencing bitcoin movement. To dig it deeper, different news are characterized by different patterns. Regulation news tend to have lagging response to bitcoin price movement since it takes time for the regulator to take legal action against bitcoin trading. On the other hand, bitcoin price tends to be driven by important regulation news published by governments or organizations in a relatively short time period. WSJ news, which focus on market news, often thrives immediately once there is a bitcoin story to talk about. Privacy security news will only grab public attention when there is a privacy security breach on cryptocurrency. The corresponding impact of privacy security news is prone to be detrimental.

8.1 Prediction Models

Comparing the three type of models implemented in this project for predicting bitcoin daily close price, unsurprisingly, LSTM outperforms Linear Regression and SVR. Since the mechanism of LSTM, it has "memory" to deal with long sequence or time series data. We started training the model by 1-day time stamp and use various combinations of input features (open, close, high, low bitcoin price, volume converted to percentage changes refer to previous day, moving average, USE/EUR ratio, number of tweets and oil price). However, the predicted results are almost the same as the previous day's actual close price, therefore, we did experiments with different choice of time stamps and input feature combinations. As shown from the model result section, the best input feature combination is open, close, high, low prices,

volume(in percentage) and 9-day moving average while setting time stamp to 10 days and run for 140 epochs. Based on these configurations, the prediction resuls of test set has 115.17 RMSE and 62.10% accuracy on predicting price rise or fall on the next day. Although the classification results of bitcoin price rise or fall have acceptable accuracy, it is hardly to be used in practical due to relatively high RMSE of exact price prediction, which means the magnitude of rise or drop in price can not be predicted.

In (Caton, 2018) the author implemented a LSTM model to classify bitcoin price movements. The bitcoin dataset used ranges from the 19th of August 2013 until the 19th of July 2016, 80/20 holdout validation strategy is used. Model is trained by daily open, close, high, low bitcoin prices, mining difficulty and hash rate. And the model achieved 52.78% accuracy by 100-day time stamp, which is lower than our LSTM model accuracy. Therefore, we conclude that adding technical factors e.g. moving average may improve prediction accuracy. Moreover, trading volume changes in percentage may also play an important role in bitcoin price prediction.

8.2 Implication and Recommendation

The implication of the project is four folded, a successful delivery of this project will grab interests from four stakeholder groups.

1.Financial institutions: finance professionals are embracing the era of AI and Big Data. Machine learning has been widely adopted for modelling many financial products and derivatives so modelling bitcoin is no exception. Our project has unique values offered for them. First, historical price and other cryptocurrency price are the determining factors for bitcoin analysis. When comparing to fundamental analysis, financial institutions ought to put more emphasis on technical analysis since historical price per se has already reflected the impact of a vast variety of external factors. In addition, our trained model is of much referential value to them because the model can make price prediction in the short term at a considerable level of accuracy.

2.Investors: For most investors, investment decision making is in good part determined by social media, hearsays and technical indicators. However, they often suffer significant financial loss due to a lack of comprehensive understanding of bitcoin insights. Therefore, our project can serve as a reliable knowledge source to them as to figure out which external factors investors should look at when making investment decisions on bitcoin. Specifically, investors should pay more attentions to social media sentiment in the case of bitcoin price surge or avalanche because press often provides meaningful insights about the market if there are big issues about bitcoin. Additionally, our model can be a valued asset for them, providing guides to facilitate their investment decision making. As a result, we expect our model can help investors to manage investment risk to an acceptable level. Overall, based on the findings of the project, investors need to make bitcoin investment decision primarily on technical indicators.

3.Education institutions: although there has been numerous papers and projects concerning with studying external factors of bitcoin price and modelling bitcoin price prediction, our project can still deliver academic value to the future research, such as factor identification, feature importance analysis, news

sentiment analysis, visualization as well as modelling. In addition, we cannot deny the inherent limitation of our project. Thus, we hope this project can be a useful reference for future research so that that all the limitations in this project can be properly tackled.

4.Policy makers: Given the fact that bitcoin is a decentralized currency, the diffusion of bitcoin has already grabbed the attention of governments in the globe. This is because the existence of bitcoin nourishes illegal crossings through informal channels especially money laundry. If decision makers from the government plan to control the bitcoin traffic, they have to get a clear perspective of bitcoin, including what external factors are likely to impact bitcoin price. Therefore, our project has its own value to offer since we provide in-depth feature importance analysis, regulation news sentiment, visualization as well as machine learning modelling to explore various avenue of bitcoin price movement pattern. Ideally, these findings can be utilized to assist government with regulation and legal control on bitcoin usage. Most importantly, the research has found that most bitcoin regulations tend to be corrective, which are introduced several months after bitcoin crisis. Corrective regulation is often the sub-optimal solution because investors and the market have already suffered heavy financial losses before legal actions are taken. Thus, having learnt the lesson of bitcoin crisis at the beginning of 2018, the government can learn valuable insights from the project and attempt to proactively introduce preventive legislations in order to prevent bitcoin crisis in the first place. If the risk of bitcoin crisis can be mitigated to an acceptable level, then there is no need to legislate corrective policies against bitcoin.

In this project, we succeed in narrowing some of the existing gaps in the field of study. To begin with, our LSTM model outperforms the counterpart trained in previous research. Second, we identify and collect 16 factors(including both foundamental factors and technical factors) to analyze the importance of each factor in relation to bitcoin price. Moreover, we manage to divide news into regulation news, market news as well as privacy and security news in order to study the impact of news sentiment in each subcategory. As a result, we yield intuitive and meaningful insights of news sentiment in each fraction.

In brief, based on the findings and implications of the project, investors are recommended to put more weights on technical factors such as historical moving average and volume when making investment decisions. In addition, investors ought to pay close attention to news sentiment in the case of bitcoin price surge or sharp decline when bitcoin news sentiment reflects the public perception of bitcoin as well as future trending. Moreover, government is suggested to proactively monitor the crpytocurrency market and introduce preventive legislation to prevent bitcoin crisis before it is too late.

9 LIMITATIONS AND FUTURE WORKS

Most of limitations are related to data collection and data analysis. In data collection, the original plan is mining twitter about bitcoin from 2015 to 2019. But the twitter API has mining constrains, the API only allows users to mine twitter up to seven days. Also, many tweets do not provide meaningful insights on bitcoin. Therefore we decide to forgo the use of news API.

In addition, we plan to mine news by using news API. The news API can be used to mine articles up to 12

months but in order to take full advantage of the API, an enterprise or premuin account is required. The developer version does not support advanced news search. Alternatively, we decide to collect news from Bitcoin magzine and develop python news scrapper to mine bitcoin news from the Wall Street Journal.

The inherent limitation in the news sentiment analysis is the limited scope of news collection. In this project, only English bitcoin news are collected. Moreover, although VADER lexicon is highly regarded in terms of performance and flexibility, its inherent limitation shall not be ignored. To begin with, sentiment analysis via VADER lexicon is a bag of word method. Under bag of word method, each word in the news is tokenized and the sentiment score for each individual token is calculated. However, it fails to consider the semantic meaning of adjacent word and phase. Moreover, VADER lexicon is reliable and accurate, but it does not guarantee a 100 percent accuracy on sentiment due to the complicated nature of news content. When performing sentiment analysis on news with arguments on both positive side and negative side, VADER lexicon may likely to derive sub-optimal sentiment score. Furthermore, news per se is intrinsically complex with subtle meanings. Undoubtedly, the opinion on a single piece of news may vary one person to another because news sentiment is subject to human perceptions and discretion. In some cases, there is no absolute positive or negative sentiment for news, it is depending on how reader understand, interpret and argue from their own perspective.

The inputs of LSTM model are collected by daily values. By experiments and analysis, using smaller time period data (hourly) as inputs may improve model performance of predicting bitcoin price. Due to limitation of access to data sources/API, we are not able to perform more accurate predictions.

The limitations mentioned above could become the starting point for feature work. To begin with, we can purchase the News API premium licence to facilitate the news collection process while taking news in other languages into consideration. Second, we can opt to use machine learning or deep learning based sentiment analysis to see if the performance of news sentiment analysis outperforms the current one. Last but not the least, we can manage to mine the digital wallet address from the blockchain website and map the addresses to the real addresses in order to find out which platform or individual is manipulating bitcoin price via engaging a huge volume of bitcoin transactions. Ideally, we can attempt to predict bitcoin price by studying the trading pattern and behaviour of these addresses

REFERENCES

Adebiyi A A, Ayo C K, A. M. O. O. S. O. (2012). Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*.

Al Shehhi, A., Oudah, M., and Aung, Z. (2014). Investigating factors behind choosing a cryptocurrency. In *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 1443–1447. IEEE.

Caton, S. M. J. R. S. (2018). Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction. *School of Computing, National College of Ireland, Dublin 1, Ireland*.

Chatfield, C. and Yar, M. (1988). Holt-winters forecasting: some practical issues. *The Statistician*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. pages 785–794.

Ciaian, P., Rajcaniova, M., and Kancs, d. (2014). The economics of bitcoin price formation. *St. Louis: Federal Reserve Bank of St Louis*.

Donges, N. (2018). The random forest algorithm. *Towards Data Science*.

Dwyer, G. P. (2015). The economics o f bitcoin and similar private digital currencies. *Journal o fFinancial Stability*.

et al, D. H. (2016). Virtual currencies and beyond: Initial considerations.

Federal Reserve Board, P. H. P. (2015). Bitcoin: Technical background and data analysis. south carolina: Createspace independent publishing platform.

Gronwald, M. (2015). The economics of bitcoins: News, supply vs demand and slumps. 15(2):157.

H. Jang, J. L. (2018). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access*.

Hayes, A. (2015). What factors give cryptocurrencies their value: An empirical analysis. *Available at SSRN 2579445*.

Hayes, A. S. (2017). Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin. *Telematics and Informatics*, 34(7):1308–1321.

I. Georgoula, D. Pournarakis, C. B. D. N. S. and Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *SSRN 2607167*.

Kaastra, I. and Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, vol. 10, no. 3*.

Kharde, V. A. and Sonawane, S. (2016). Sentiment analysis of twitter data: A survey of techniques. *International Journal of Computer Applications*.

Kshirsagar G, Chandel M, K. S. A. R. (2018). Stock market prediction using artificial neural networks. international journal of advanced research in computer engineering and technology.

Lewis, N. (2017). What is the fundamental value of bitcoin? *Forbes*.

Li, X. and Wang, C. A. (2017). The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin. *Decision Support Systems*, 95:49–60.

M. Brie‘re, K. O. and Szafarz, A. (2013). Virtual currency, tangible return: Portfolio diversification with bitcoins. *Tangible Return: Portfolio Diversification with Bitcoins*.

Majaski, C. (2019). The difference between fundamental vs. technical analysis? *Investopedia*.

Mallqui, D. C. and Fernandes, R. A. (2019). Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, 75:596–606.

Mandjee, T. (2015). Bitcoin, its legal classification and its regulatory framework. *Journal of Business Securities Law*, 15(2):157.

Nakamoto, S. (2008). “bitcoin: A peer-to-peer electronic cash system“.

Navickas, I. B. (2018). Predicting bitcoin price using machine learning.

Nguyen, T. (2018). Illustrated guide to lstm’s and gru’s: A step by step explanation.

Nian, L. P. and Chuen, D. L. (2015). Introduction to bitcoin. *Handbook of Digital Currency,*.

Numnonda, T. P. (2018). Machine learning models comparison for bitcoin price prediction. *King Mongkut’s Institute of Technology Ladkrabang*.

Phaladisailoed, T. and Numnonda, T. (2018). Machine learning models comparison for bitcoin price prediction. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 506–511. IEEE.

Radityo, A., Munajat, Q., and Budi, I. (2017). Prediction of bitcoin exchange rate to american dollar using artificial neural network methods. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 433–438. IEEE.

Rechenthin, M. D. (2014). Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction.

S. Vassiliadis, P. Papadopoulos, M. R. T. K. (2017). Bitcoin value analysis based on cross-correlations. *IEEE Access*.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero.

Sukamulja, S. and Sikora, C. O. (2018). The new era of financial innovation: The determinants of bitcoin’s price. *Journal of Indonesian Economy and Business*, 33(1):46–64.

Tonidandel, S. and LeBreton, J. M. (2015). Rwa web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology*, 30(2):207–216.

Wah B W, Q. M. L. (2006). Constrained formulations and algorithms for predicting stock prices by recurrent fir neural networks. international journal of information technology and decision making.

White, H. (1988). Economic prediction using neural networks: The case of ibm daily stock returns in neural networks. *IEEE International Conference on. IEEE*.

Y. Peng, P.H.M. Albuquerque, J. d. S. A. P. M. M. (2018). The best of two worlds: forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Syst*.