Bitcoin price prediction based on sentiment analysis and LSTM

Chenfeiyu Wen[1,6], Xiangting Wu[2,7], Chuyue Shen[3,8], Zifei Huang[4,9] and Peiqi Cai[5,10]

[1]Computer Science and Technology, Central South University, Hunan, 410083, China
[2]The Experimental High School Attached to Beijing Normal University, Beijing, 100032, China
[3]School of Science, Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China
[4]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, 43212, The United States of America
[5]Zhejiang University of Illinois at Urbana-Champaign Institute, Zhejiang University, Haining, 314499, China

[6]1076514781@qq.com
[7]tinawu2018@sina.com
[8]chuyueshen615@gmail.com
[9]huang.4078@osu.edu
[10]peiqic3@illinois.edu

Abstract. As cryptocurrencies become widely accepted due to technical improvements, reliable approaches to capture their future price movements of them become critical. This study mainly combines weighted sentiment analysis results from social media-related comments and financial news headlines with a stacked LSTM model to predict second-day Bitcoin price evolution. This study also compared our results and the results produced by MLP, RF, and SVM after feeding the sentiment analysis results.

Keywords: Bitcoin, LSTM, sentiment analysis, price prediction.

1. Introduction

On November 1, 2008, Satoshi Nakamoto (pseudonym) published *Bitcoin: A Peer-to-Peer Electronic Cash System*, in which he stated his pioneering vision of decentralization. He developed and open-sourced the Bitcoin system, and mined the first block of the Bitcoin blockchain on January 3, 2009. The price of Bitcoin rose from 10,000 coins for two pizzas (May 22, 2010) to 10,000$ for one coin (February 10, 2020).

Bitcoin is a decentralized system that attempts to overcome the weaknesses of credit currencies and precious-metal-based convertible paper money. It is not governed by central authorities such as governments or banks, and users can allodially change the transaction fee. The technologies involved in Bitcoin, such as cryptography, distributed ledger, and decentralization, are collectively known as blockchain technology, which has numerous applications in finance, the Internet of Things, insurance, et cetera.

Bitcoin is also the first platform in which humans relied entirely on technology to protect their property instead of the rule of man or law. The significance behind this is immeasurable, and Bitcoin will likely set off a monetary revolution in the future. However, concerns still exist, and with more people involved in Bitcoin with a speculative mindset, predicting the market value of Bitcoin is necessary.

L. Kristoufek examined the dynamics between the Bitcoin price fluctuations and several possible sources through wavelet coherence to find the main drivers of the Bitcoin price [1]. Overall, financial factors like usage in trade, money supply, and price level play a critical role in the long term but are prone to bubbles and busts in the short run; public interest in Bitcoin not only impacts the long-term fluctuations but also drives prices further up or down according to the existing trend.

Social media posts can reflect the public interest and the masses' perspectives. Similarly, financial reviews about Bitcoins will influence vast scales of people through expressing their assessments and predictions. Previous researchers have proved sentiment as a predictive factor in an asset pricing framework. Tetlock discovered that sentiment polarities, especially pessimism, predict pressure on market prices [2]. Karalevicius et al. found that expert media performs the role of a good short-term predictor of future Bitcoin price movements [3].

This project aims to use the Long Short-Term Memory (LSTM) network combined with sentiment analysis to predict Bitcoin price. Since Bitcoin is decentralized and is highly related to public interests and perspectives, we analyzed the sentiments of the users' comments with "Bitcoin" or "BTC" on Twitter and related financial headlines. Then, we combined the sentiment results with recurrent neural networks to predict. Finally, we evaluated their accuracy and other metrics, and assessed their improvements or retrogress compared with traditional machine learning algorithms.

Here are the main outcomes of our paper:

1. We processed the data from the datasets and analyzed the sentiment polarities through natural language processing.
2. We combined the sentiment polarity results with Long Short-Term Memory (LSTM) and adjusted the parameters to reach better performance in predicting the second-day Bitcoin price.
3. We implemented Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) to predict the second-day Bitcoin price and compare the results with our LSTM predictions.
4. We evaluated the models' results through Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) and examined the possible reasons.

The rest of the paper is organized as follows: Section II presents the related work and backgrounds. Section III introduces the data. Section IV clarifies the methodologies. Section V assesses the results, and finally, Section VI concludes the paper.

2. Related work and backgrounds

Bitcoin price forecasting is analogous to other time-series analyses. It can be defined as predicting the future trend of financial products by analyzing historical prices and volume. Numerous methods could be used to make accurate predictions.

2.1. Statistic models

The various kinds of statistical models are the Auto-Regressive model (AR), the Moving Average model (MA), the Auto Regressive Moving Average model (ARMA), and the Auto-Regressive Integrated Moving Average model (ARIMA), which are all generalized linear models. The ARMA model is a combination of AR and MA with different coefficients. Based on predefined rules, the output of ARMA is a univariate linear combination of historical data. The ARIMA model is a variation adjusted of ARMA for analyzing time-series data, which uses the adjacent difference of historical data to get a stationary input. Ariyo et al. used ARIMA to predict short-term stock returns [4]. They compared different models by modifying the parameters of the autoregressive and moving average. Experiments showed perfect accuracy of prediction. The model has a better performance compared to SVM in training comparatively stationary data [5]. However, it is inefficient in processing underlying dynamic information in Bitcoin close data [6]. They suggested that the class imbalance in the predictive portion of the ARIMA forecast, which is also the main drawback of this model, caused the consecutive increase in predicted price.

2.2. Machine learning models

The machine learning models involve methods like Support Vector Machine (SVM) and Deep Neural Networks. SVM is a supervised learning algorithm that seeks to minimize the generalization error with less variation [7]. It has great performance in time series forecasting with regard to its resistance to overfitting. Lin et al. used a correlation-based SVM filter to select a subset of financial indexes as input features and applied quasi-linear SVM to predict the movement direction of the stock market [8]. LeCun et al. combined ARIMA with SVM to capture the latent dynamics existing in financial data and improved overall forecasting performance combined with pure ARIMA [9]. Deep Neural Networks can approximate arbitrary functions. Different neural network structures have been proposed to fit various data processing applications. These include Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) [9]. They have been applied in various areas like computer vision, natural language processing, time series analysis, etc. Mittal et al. investigated training an RNN model to make Bitcoin price predictions and compared the results with the ARIMA model [10]. Experiments indicate that designed to remember historical data, RNN significantly outperforms the linear model. Since it is more capable of long-term dynamic change in the dataset. Additionally, the introduction of the forgetting gate enables LSTM to incorporate the factor of historical data into forecasting, contributing to its ability to recognize longer-term dependencies. Nelson et al. used LSTM to predict the future stock price of several companies [11]. Experiments illustrate that LSTM generally outperforms MultiLayer Perceptron (MLP) and Random Forest (RF) with better accuracy and lower error. Ojo et al. investigated the usage of a stacked LSTM neural network, which is composed of multiple layers of LSTM neurons [12]. The MSE and MAE indicators suggested the advantage of a multi-layer of LSTM neurons in processing dynamic stock price information.

2.3. Sentiment analysis

The purpose of Sentiment Analysis (SA) is to find opinions in texts and then identify and classify their polarity. The result of SA can help people make decisions on stock markets, news articles, or political debates [13]. Microblogging websites are becoming a popular source of information [14]. Thus, discovering an overall sentiment is practical and feasible. Twitter is one of the more popular social networking platforms where users can comment on any topic. Therefore popular sentiment is hidden in tens of millions of daily comments.

Many studies have indicated that mass sentiment can be extracted from online comments. It is shown that the Dow Jones Industrial Average (DJIA) can be predicted by daily Twitter feeds and the accuracy is 87.6% which reduces the Mean Average Percentage Error by more than 6% [15]. The baseline model performed worse than Naive Bayes and Maximum Entropy which suggests that the amount of features is related to the model's performance, due to the Baseline model having two fewer features than the other two [16]. When a dataset contains a large amount of data, LSTM models perform better [17].

In addition to some widely used algorithms, various improved models have been proposed in recent years to solve SA problems. Some people combine the advantages of different algorithms and propose new models to combine their advantages. It is known that deep convolutional networks and recurrent networks are excellent models, but have different strengths and weaknesses. Behera et al. proposed CO-LSTM models, which have great compatibility with different fields and adaptability for processing mass social data [18]. Others are working on refining the way sentiments are extracted. Huang et al. argued that traditional deep learning models perform feature extractions of sentences or words, so the emotions implied in the context cannot be analyzed. They integrate emotional intelligence and attention mechanisms to improve LSTM models [19].

2.4. Combination

Bitcoin price prediction can be based on the analysis of the stock market or other cryptocurrency price predictions. Tiwari et al. pointed out that predicting stock prices becomes a big challenge due to the nonlinearity of stock indices [20]. Lamon et al. analyzed the link between sentiment in the news and comments on social media and the actual price of cryptocurrencies (including Bitcoin, Litecoin, and Ethereum) on a given day in the future, and correctly predicted the largest percentage of price fluctuations [21]. Although sentiment-based predictions are feasible, he also suggests that there is still a lot of room for improvement in the model.

Therefore, more research combines machine learning models and sentiment analysis to predict future price trends by analyzing past price datasets and socially collected sentiment datasets on stock prices. Keyan et al. pointed out that the social interactions from social media platforms are closely related to stock price fluctuations, and the accuracy of the LSTM model is slightly higher than that of RNN and GRU (Gated recurrent unit) [22]. Two other studies from Jin et al. and Thormann et al. also, focus on predicting stock price by both sentiment analysis and LSTM [23,24]. In Jin et al.'s study, his team improved the LSTM by EMD (empirical mode decomposition) to turn a complex stock price sequence into a simpler and more predictive sequence. The results of both experiments show that the LSTM model incorporating sentiment analysis is more advantageous in stock price prediction than the other models. The study by Thormann et al. guides our group on using twitter-based variables for forecasting purposes. Finally, the study by Raju et al. applied sentiment analysis and supervised learning methods to analyze the correlation between Bitcoin price movements and sentiment in tweets and compared several algorithms that deal with time series [25].

Although data from Twitter and Reddit alone may not be comprehensive enough to represent everyone, adding sentiment analysis and considering other factors, the prediction can still become more convincing. The results show that LSTM analyzes the time series data of past Bitcoin prices more efficiently than ARIMA, and LSTM can store long-term dependencies better. Also, the dangers associated with the cryptocurrency space can be better assessed through machine learning.

3. Data

3.1. Bitcoin price

The Bitcoin and united states dollar (BTC-USD) price data is sourced from Yahoo Finance, with a time spanning from the 1st of January 2017 to the 31st of December 2018. The data are daily recorded values of the opening, closing, highest, and lowest price of Bitcoins, and the total volume transacted.

3.2. Top financial news titles

The top financial news titles are from a data set called Bitcoin Price Prediction based on News Headlines from user Mohd Abdul Azeem on Kaggle, collected from the Bitcoin News (BTCTN) Twitter user account. This data set has a total of 33631 available headlines from the 1st of July 2015 to the 12th of June 2021 (Table 1).

Table 1. Example Bitcoin financial headlines.

Date	Text
4-1-2017	The Ross Ulbricht Legal Defense ’Free Ross’ Was Hacked
10-8-2017	Australian Primary School Students Explore Bitcoin Bitcoin Cash Gains More Support
16-11-2018	A Brief Introduction to Voluntaryism for Crypto Neophytes Free Keene Activists Launch Bitcoin Embassy New Hampshire

3.3. Bitcoin tweets

We adopted another data set on Kaggle, called Bitcoin tweets-16M tweets from user Alex. This data amount to a total of 18,810,521 tweets containing "Bitcoin" or "BTC" between the 1st of January, 2016 and the 29th of March, 2019 (Table 2).

Table 2. Example Bitcoin tweets including ”Bitcoin” Or ”BTC”.

Timestamp	Likes	Text	Score
2017-01-01	4	Dear Bitcoin you are getting out of reach to those who need you the most and who you set out to free altcoins	0.387
2017-12-08	1	BTC investing is more speculative since there is very little way to judge it s intrinsic value Scarcity helps but since it no longer really functions as a useful transaction medium with the increasing fees it s true utility is harder to understand than USD	0.023
2018-04-05	0	If a network is not secure how valuable is it	-0.556

3.4. Preprocessing

Natural language, as a human resource, tends to follow the intrinsic nature of the randomness of its creator. Data preprocessing is to convert texts from human language into a machine-readable format, minimizing the effect of randomness. This procedure is done by text normalization, in which researchers usually perform the following steps: conversing letter-case, deleting unnecessary information (numbers, punctuations, accents, et cetera), removing stop words, omitting the nulls and voids, extracting word roots, and so on.

We preprocessed the data accordingly as follows. For all three datasets, we first read them into Data Frame due to their compatibility with different formats, such as HDF and excel, and their relatively fast storage speed. We then removed all data outside of January 1, 2017, to December 31, 2018, because only within this time interval is the quality of the Bitcoin tweets data guaranteed. Then, we deleted all the data of all the uniform resource locators (URL) Bitcoin tweets and, at last, omitted all the non-English characters in both the Bitcoin tweets and the financial headlines about Bitcoin.

4. Methodology

4.1. Problem formation and considerations

In general terms, the problem we are solving is to forecast the Bitcoin price using historical information which contains historical prices, tweets, comments, and news. Based on the potential characteristics of the Bitcoin market, the problems of Bitcoin price prediction can be transformed into considerations of our model as follows.

C1 - Sentiment Analysis. The value of Bitcoin is based on public consensus. The sentiments extracted from news titles and comments from social media are important in forecasting.

C2 - Capturing Potential Dynamic Information. The decentralization of Bitcoin contributes to its volatility. Just using a single prior price of Bitcoin is not sufficient, as we need a model that incorporates historical information.

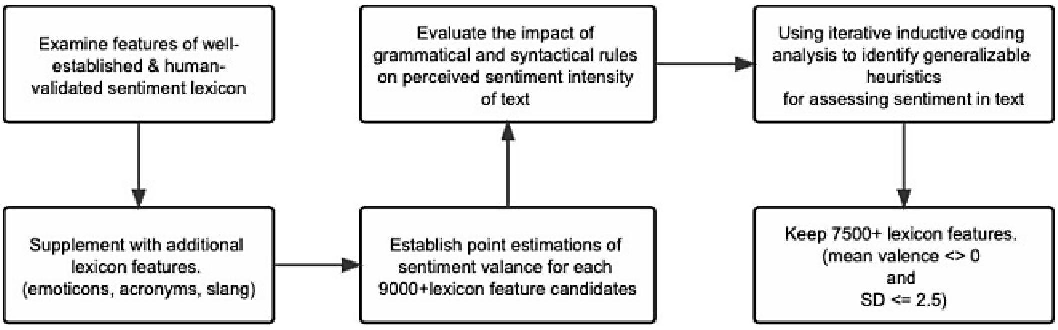
C3 - Different Weights for Information. The sentiments extracted from different platforms may have different weights. Even the same comment from the same person with a different count of likes and retweets may represent different influences on the Bitcoin market.

4.2. Sentiment analysis

Natural Language Toolkit (NLTK) is a suite of open-source program modules that include symbolic and statistical natural language processing methods [26]. Vader is a popular rule-based sentiment analysis model, and the model is designed for social media content [27]. This model is prominent in treating tweet comments and news headlines regarding sentiment extraction (C1).

Unlike machine learning methods such as Naive Bayes, and Support Vector Machines, Vader has an artificially constructed sentiment lexicon, which lowers the amount of training data needed and makes it fast in processing.

Vader's sentiment lexicon is based on some well-established corpus, which then introduces a large number of Internet terms on this basis, such as emoticons, slang, acronyms, and initialisms. Through this process (Figure 1), the word library is expanded into a lexical feature candidate set of more than 9000 words. Then, through artificial polarity labeling and sentiment intensity scoring on Amazon Mechanical Turk (AMT), a sentiment dictionary containing more than 7500 words with sentiment polarity and sentiment intensity is obtained.


Examine features of well-established & human-validated sentiment lexicon

Supplement with additional lexicon features. (emoticons, acronyms, slang)

Establish point estimations of sentiment valance for each 9000+lexicon feature candidates

Evaluate the impact of grammatical and syntactical rules on perceived sentiment intensity of text

Using iterative inductive coding analysis to identify generalizable heuristics for assessing sentiment in text

Keep 7500+ lexicon features. (mean valence <> 0 and SD <= 2.5)

Figure 1. Overview of lexicon approaching.

Going further, researchers manually scored 800 tweets for emotional intensity. They then used a method similar to the Grounded Theory approach to identify features that affect the polarity and intensity of the text [28]. This results in five heuristics that are more astute than typical bag-of-words models for discovering the influence of features such as characters, capitalization, etc. in context on sentiment levels.

By introducing Vader, we can extract the features in the text, and then perform polarity classification and sentiment analysis on the features, so that the sentiment of each text in the dataset can be quantified.

4.3. LSTM

In considering the potential characteristic of the Bitcoin market (C2), we use a stacked LSTM neural network to make the prediction. The cell of the LSTM neuron consists of an input gate, an output gate and a forget gate. Figure 2 illustrates the structure of an LSTM network with its 3 gates.

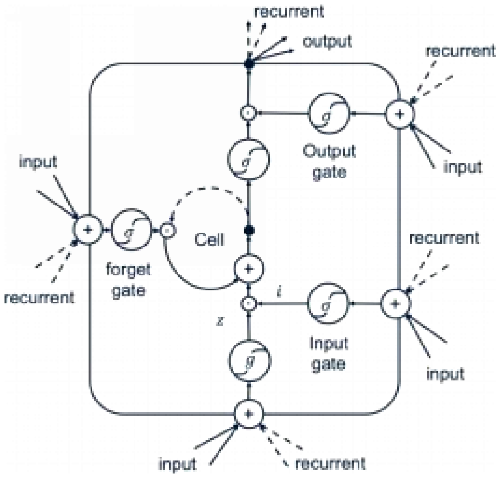


Figure 2. LSTM neuron.

The introduction of forget gate enables the LSTM neuron to remember historical input by adding them to the memory cell. The output of the gates can be represented as follows:

$$c^t = z^f \odot c^{t-1} + z^i \odot z \tag{1}$$

$$h^t = z^o \odot tanh(c^t) \tag{2}$$

$$yt = \sigma(W'h^t) \tag{3}$$

$\odot$ is Hadamard Product (or element-wise product),

C^t represents the information stored in the memory,

Z^f represents the controller of the forget gate,

y^t stands for the output of the model, similar to RNN, it is also from h^t.

Training the LSTM network with whole records of historical Bitcoin prices enables the model to remember latent connections between the daily information given by the training set, and make the prediction with a long-term perspective.

The stacked LSTM neural network provides a deeper model for learning. It is composed of multiple hidden layers of LSTMs, enabling the hidden layer to learn the inner connections between the sentiments and the historical prices, contributing to the high accuracy of prediction.

4.4. Sentiments with LSTM neural network

To make sentiments from online social media an auxiliary part of Bitcoin price prediction, we combined sentiment analysis with LSTM. It is important to input the historical Bitcoin prices and sentiment analysis results into the stacked LSTM neural network. A value of compound can be extracted from each tweet message, forming a large feature matrix of the result of sentiment analysis. We regularized the feature matrix of SA, aggregated the compound each day into a single numeric value, and fed them into the input layer of the neural network as a second dimension of the input matrix. Based on the count of replies, likes and retweets, we give different weights (C3) to each comment. The compound of each day can be represented as follows:

$$s^t = \frac{1}{n}\sum_{i=0}^{n} lg\left(c^l + 1\right) \times lg(c^{rt} + 1) \times lg(c^{rp} + 1) \times c^c \tag{4}$$

where:

s^t stands for the sentiment feature of tweets of day t,

n stands for the count of comments of day t,

c^l, c^{rt}, c^{rp} and c^c stands for the count of likes, retweets, replies, and the compound from sentiment analysis.

As formula (2) suggested, we used the log function of each tweet feature instead of a linear one to exclude outlier interference. The difference between the weight and the log of weight is illustrated in Figure 3. Specifically, if a single tweet receives a significant amount more replies, likes, or retweets than others, the sentiment feature this day may explode and mute the sentiments from other days, as (A) of Figure 3 suggested. Except for the two days in the center of the figure shows relatively high compound values of sentiments, which are both close to 1, the values extracted from other days are muted significantly, around 0. Additionally, the relationship between the popularity of a single comment and its influence is not linear. As the number of follows, likes, and retweets of a tweet grows, the recommendation algorithm of Twitter shows it to more people, making its statistic data explode. We used the log function to restore the data to the normal range to show the fluctuation of sentiments precisely, as (B) of Figure 3. shows, which is more rational to be fed into the subsequent neural network.

We combined vector ***s***, which is a combination of s^t of each day, with historical price $\boldsymbol{p}$ as the second dimension of the input matrix of the LSTM neural network. Two hidden LSTM layers were chosen to capture the potential connections between the historical data and future prices, with 128 and 64 neurons each. For a time series task, a fully connected layer is enough to find nonlinear relationships among the data with 64 nodes. The output layer is a fully connected layer, enabling the model to take all the output of the hidden layer into account. It is composed of a single neuron to output the prediction of the next day Bitcoin price log return.

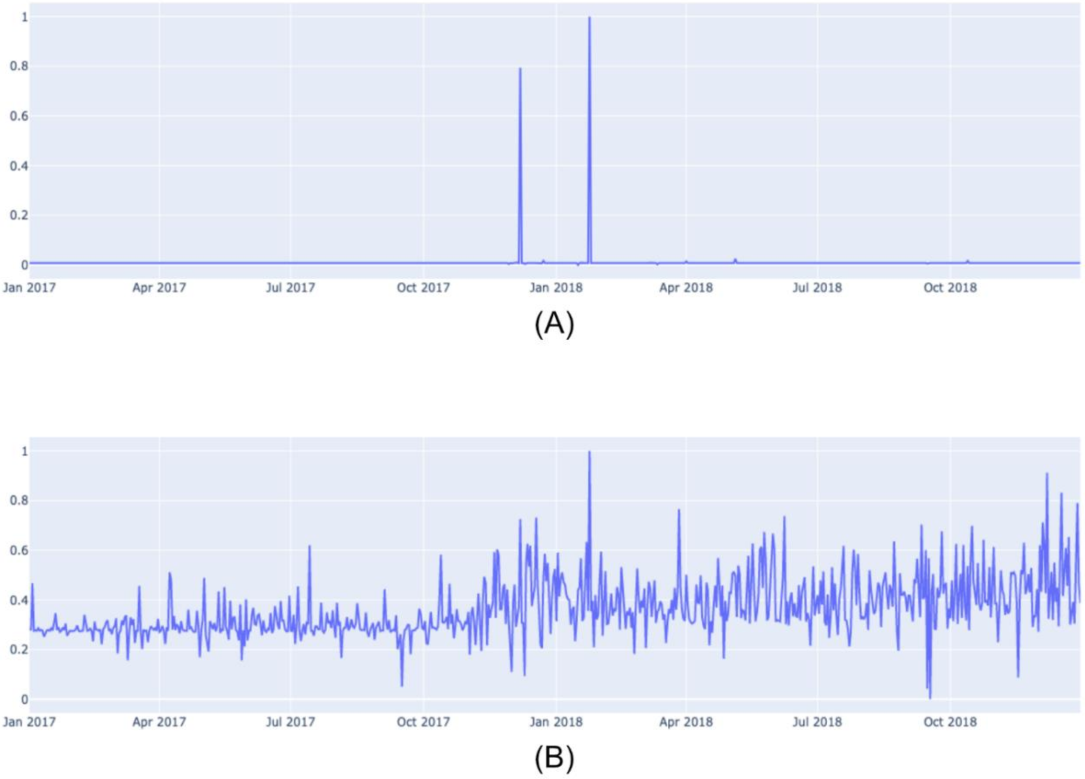


Figure 3. A comparison of different functions of weight. (A) shows the fluctuation of sentiments with a linear function of weight, (B) shows the fluctuation of sentiments with the log10 function of weight, both from Jan 1, 2017, to Dec 31, 2018.

4.5. MLP, RF, and SVM

To make comparisons, we also fed our data into MLP, RF, and SVM models with appropriate coefficients. In the MLP model, we used 100 neurons in the hidden layer, ReLU (Rectified Linear Activation Function) as the activation function, and Adam boost as a solver for weight optimization with the learning rate set to 0.001. In the RF model, we used 100 decision stamps with 1 max depth to control overfitting. In the SVM model, we chose Radial Basis Function as the SVM kernel, set the regularization parameter to 1.0, and set $\mathcal{E}$ to 0.1 to prevent the prediction return not too far from the true one. The input data is the same as the data fed into ours.

4.6. Evaluation

We predicted the log return of Bitcoin price through time-series data sets and sentiments, so methods for our group model performance assessment focus on the difference between the actual return and their predicted values. Two Methods can be applied to evaluate the accuracy of forecasting. The first is the Mean Absolute Error (MAE). MAE measures (Formula 3) the average difference between predictive data and true value and does not consider the direction of the error.

$$MAE = \frac{\sum_{i=1}^{n}|\widehat{y_i}-y_i|}{n} \tag{5}$$

where y_i is a predicted value, x_i is a true value, and n is the number of samples.

The second method is Root Mean Squared Error (RMSE).In this equation (Formula 4), S_i is a predicted value, O_i isa true value, and n is the number of samples. Although RMSE also measures the average difference between the predicted value and the observation value throughout the prediction process, it will easily detect large errors, because the square is before the root operation, and the larger difference in the predictionis given greater weight.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{6}$$

These evaluation methods can be used simultaneously to diagnose errors in a set of prediction log returns, and the value of the RMSE will always be greater than the MAE value. The difference between RMSE and MAE represents the change between error, the larger the difference between RMSE and MAE, the greater the variance of each individual error. If RMSE = MAE, the change between errors is 0, that is, each group of errors has the same size. The ranges of MAE and RMSE are both from 0 to positive infinity, and the model's performance is better when the results of RMSE and MAE are closer to 0.

To measure the accuracy of each single prediction value, we will determine whether the model successfully predicts the rise or fall in the second-day price. We will compare the signature of the predicted log return with the true value and calculate the percentage of the same signs, which suggests for the correct prediction. The larger the percentage is, the better model we made for investors to refer to.

5. Results

We chose the last 20 percent of the whole data set as the test set. We feed the sentiment compounds and the price of the day before into our network to get the predicted log return $\hat{y}$ and compare it with the true values. To make better visualizations, we used the predicted log return to calculate the bitcoin price and plot it with the true price on the same chart. Fig 4 shows the calculated predicted Bitcoin price of several models versus real price of the test set. Several indicators are used to analyze the performance of the models. We do not draw the result of MLP because of the extreme bias of predicted price. In this chart, it is clear that the trend predicted by RF and SVM are both far from true value, showing their inability to capture the dynamic information of historical data.

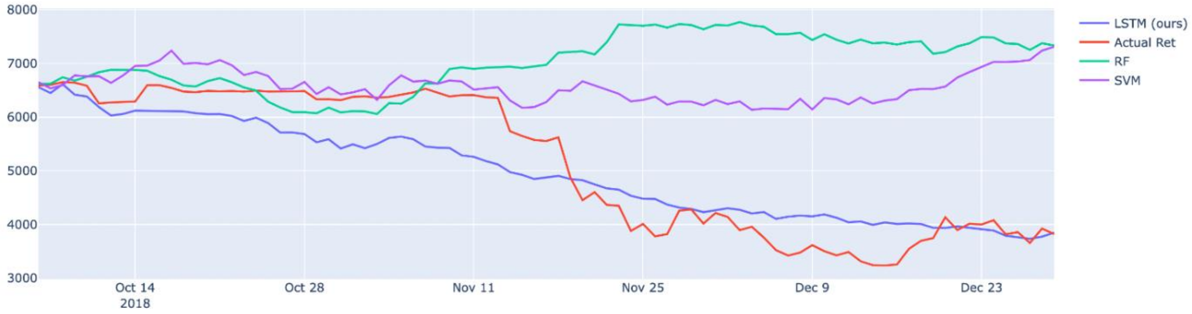


Figure 4. Calculated predicted price vs actual price.

The figure shows the comparison of 3 models on the predicted price and the true price of bitcoin.

5.1. *MAE and RMSE*

We implemented both MAE and RMSE on the test set to evaluate (Table 3). The result of the MAE is 148.1383. Based on the scale of values in the test data set, the MAE value shows that the LSTM predictions are fairly accurate. The average error of predictions is about 148 dollars out of several thousand dollars. The RMSE, which gives a larger weight to the greater error of prediction, generates a result of 213.6116. With this result, we can conclude that LSTM performs decently in predicting Bitcoin price since there is no unacceptably large error. Based on the difference between MAE and RMSE, the error of our predictions varies within an acceptable scale.

Table 3. Evaluation of each model.

Model	MAE	RMSE
SVM	0.0331	0.421
RF	0.0331	0.0445
MLP	0.0549	0.0671
LSTM(OURS)	0.0285	0.0396

5.2. Prediction of bitcoin directional movements

We also determined the percentage of accurate forecasts of the directional moves of Bitcoin. We compared the sign of (next day predicted price - today predicted price) and the sign of (next day real price - today real price). If two signs are the same, the prediction can be considered accurate. There is a total of 74 out of 143 days with accurate predictions. However, to clarify, the predictions with the opposite sign as the real value do not mean two values are not close to each other. Those could still be accurate predictions based on the price value rather than gain or drop.

6. Conclusion

This paper presents a Bitcoin prediction model that uses recurrent neural networks represented by LSTM networks, and sentiment analysis of social media and news headlines by recurrent neural networks for integrated weighted prediction, which has higher check-all and check-accuracy rates than traditional autoregressive methods. The main contributions of this study are as follows. First, we use Stacked LSTM to capture information about Bitcoin price dynamics. The upper LSTM structure in the Stacked LSTM model outputs a sequence rather than a single value to the next LSTM structure. This stacking of LSTM layers has a much greater effect than in the past when memory cells were added inside LSTM cells. Second, we set more reasonable weights for natural language processing such as sentiment analysis. Our prediction behavior is more scientific compared to the past using only the LSTM network or only sentiment analysis prediction. Finally, we take a logarithmic approach to the results derived from sentiment analysis when dealing with outliers. Because we found that some of the outliers were very different from other values, and when we took logarithms of them, we found that the results would be much more accurate than before.

Though we forecasted the price of Bitcoin in one day with little error, under most circumstances we might need to have foresight farther than one day. We considered the feasibility of iterating with our output, but the problem is we cannot obtain sentimental data from the future for our model. Also, in terms of datasets, we have only used texts on Twitter, while many platforms besides Twitter also have similar valuable data. If we can expand the scope of data collection, it may make our sentiment analysis results more representative. In addition, the results of the Vader model can be processed using different algorithms, which may have provided a better reflection of the impact of sentiment analysis on Bitcoin. In the end, we only used one model, LSTM. Using other models, or combining different models, may yield better results.

Acknowledgment

Chenfeiyu Wen, Xiangting Wu, Chuyue Shen, Zifei Huang, Peiqi Cai contributed equally to this work and should be considered co-first authors.

References

[1] Ladislav Kristoufek. “BitCoin meets Google Trends andWikipedia: Quantifying the relationship between phenomena of the Internet era”. en. In: *Scientific Reports* 3.1 (Dec. 2013), p. 3415. ISSN: 2045-2322. DOI: 10.1038/srep03415. URL: http://www.nature.com/articles/srep03415 (visited on 03/12/2022).

[2] Paul C. Tetlock. “Giving Content to Investor Sentiment:The Role of Media in the Stock Market”. en. In: *The Journal of Finance* 62.3 (June 2007), pp. 1139–1168. ISSN: 00221082. DOI: 10.1111/j.1540-6261.2007.01232.x. URL:

https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2007.01232.x (visited on 03/12/2022).

[3] Karalevicius, V., Degrande, N. and De Weerdt, J. (2018), "Using sentiment analysis to predict interday Bitcoin price movements", Journal of Risk Finance, Vol. 19 No. 1, pp. 56-75. https://doi.org/10.1108/JRF-06-2017-0092

[4] Adebiyi A. Ariyo, Adewumi O. Adewumi, and Charles K. Ayo. "Stock Price Prediction Using the ARIMAModel". In: *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. Cambridge, United Kingdom: IEEE, Mar. 2014, pp. 106–112. ISBN: 978-1-4799-4922-9 978-1-4799- 4923-6. DOI:10.1109/UKSim.2014.67. URL: http://ieeexplore.ieee.org/document/7046047/ (visited on03/12/2022).

[5] Bhawna Panwar et al. "Stock Market Prediction Using Linear Regression and SVM". In: *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. Greater Noida,India: IEEE, Mar. 2021, pp. 629–631. ISBN: 978-1- 72817-741-0. DOI: 10.1109/ICACITE51222.2021. 9404733. URL:https://ieeexplore.ieee.org/document/9404733/ (visited on 03/12/2022).

[6] Sean McNally, Jason Roche, and Simon Caton. "Predicting the Price of Bitcoin Using Machine Learning". In: *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*. Cambridge: IEEE, Mar. 2018, pp. 339–343. ISBN: 978-1-5386-4975-6. DOI: 10.1109/PDP2018.2018.00060. URL: https://ieeexplore.ieee.org/document/8374483/ (visited on 03/12/2022).

[7] Vladimir Naumovich Vapnik. *The nature of statistical learning theory*. 2nd ed. Statistics for engineering and information science. New York: Springer, 2000. ISBN: 978-0-387-98780-4.

[8] Yuling Lin, Haixiang Guo, and Jinglu Hu. "An SVM- based approach for stock market trend prediction". In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. Dallas, TX, USA: IEEE, Aug. 2013,pp. 1–7. ISBN: 978-1-4673-6129-3 978-1-4673-6128-6. DOI: 10.1109/IJCNN.2013.6706743. URL: http://ieeexplore.ieee.org/document/6706743/ (visited on03/12/2022).

[9] Ping-Feng Pai and Chih-Sheng Lin. "A hybrid ARIMA and support vector machines model in stock price forecasting". en. In: *Omega* 33.6 (Dec. 2005), pp. 497–505.ISSN: 03050483. DOI: 10.1016/j.omega.2004.07.024.URL: https://linkinghub.elsevier.com/retrieve/pii/S0305048304001082 (visited on 03/12/2022).

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton."Deep learning". en. In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539. URL: http://www.nature.com/articles/nature14539 (visited on 03/12/2022).

[11] Aditi Mittal et al. "Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data". In: *2019 Twelfth International Conference on Contemporary Computing (IC3)*. Noida, India: IEEE, Aug. 2019, pp. 1–6. ISBN: 978-1-72813-591-5. DOI: 10.1109/IC3.2019.8844899. URL: https://ieeexplore.ieee.org/document/8844899/ (visited on 03/12/2022).

[12] David M. Q. Nelson, Adriano C. M. Pereira, and Renato A. de Oliveira. "Stock market's price movementprediction with LSTM neural networks". In: *2017 International Joint Conference on Neural Networks (IJCNN)*.Anchorage, AK, USA: IEEE, May 2017, pp. 1419–1426. ISBN: 978-1-5090-6182-2. DOI: 10.1109/IJCNN.2017.7966019. URL: http://sieeexplore.ieee.org/document/7966019/ (visited on 03/12/2022).

[13] Samuel Olusegun Ojo et al. "Stock Market Behaviour Prediction using Stacked LSTM Networks". In: 2019 *International Multidisciplinary Information Technologyand Engineering Conference (IMITEC)*. Vanderbijlpark,South Africa: IEEE, Nov. 2019, pp. 1–5. ISBN: 978- 1-72810-040-1. DOI: 10.1109/IMITEC45504.2019.9015840. URL: https://ieeexplore.ieee.org/document/9015840/ (visited on 03/12/2022).

[14] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey". en. In: *Ain Shams Engineering Journal* 5.4 (Dec. 2014), pp. 1093–

1113. ISSN: 20904479. DOI: 10.1016/j.asej.2014.04.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S2090447914000550 (visited on 03/12/2022).

[15] Apoorv Agarwal et al. “Sentiment Analysis of Twitter Data”. en. In: (), p. 9.

[16] Johan Bollen, Huina Mao, and Xiaojun Zeng. “Twitter mood predicts the stock market”. en. In: *Journalof Computational Science* 2.1 (Mar. 2011), pp. 1–8. ISSN: 18777503. DOI: 10.1016/j.jocs.2010.12.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S187775031100007X (visited on 03/12/2022).

[17] Shaunak Joshi and Deepali Deshpande. “Twitter Sentiment Analysis System”. In: (2018). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1807.07752. URL: https://arxiv.org/abs/1807.07752 (visited on 03/12/2022).

[18] Dr. G. S. N. Murthy et al. “Text based Sentiment Analysis using LSTM”. en. In: *International Journal of Engineering Research and* V9.05 (May 2020), IJERTV9IS050290. ISSN: 2278-0181. DOI: 10.17577/IJERTV9IS050290. URL: https://www.ijert.org/text-based-sentiment-analysis-using-lstm (visited on 03/12/2022).

[19] Faliang Huang et al. “Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis”. In: *IEEETransactions on Neural Networks and Learning Systems*(2021), pp. 1–14. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2021.3056664. URL: https://ieeexplore.ieee.org/document/9358000/ (visited on03/12/2022).

[20] Sachin Tiwari. “A Survey on LSTM-based Stock Market Prediction”. en. In: (), p. 8.

[21] Connor Lamon, Eric Nielsen, and Eric Redondo. “CRYPTOCURRENCY PRICE PREDICTION USINGNEWS AND SOCIAL MEDIA SENTIMENT”. en. In: (), p. 1.

[22] Keyan Liu, Jianan Zhou, and Dayong Dong. “Improvingstock price prediction using the long short-term memory model combined with online social networks”. en. In: *Journal of Behavioral and Experimental Finance* 30 (June 2021), p. 100507. ISSN: 22146350. DOI: 10.1016/j.jbef.2021.100507. URL: https://linkinghub.elsevier.com/retrieve/pii/S2214635021000514 (visited on 03/12/2022).

[23] Zhigang Jin, Yang Yang, and Yuhong Liu. “Stock clos-ing price prediction based on sentiment analysis and LSTM”. en. In: *Neural Computing and Applications*32.13 (July 2020), pp. 9713–9729. ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-019-04504-2. URL: http://link.springer.com/10.1007/s00521-019-04504-2 (visited on 03/12/2022).

[24] Marah-Lisanne Thormann et al. “Stock Price Predic- tions with LSTM Neural Networks and Twitter Sentiment”. In: *Statistics, Optimization & Information Computing* 9.2 (May 2021), pp. 268–287. ISSN: 2310-5070, 2311-004X. DOI: 10.19139/soic-2310-5070-1202. URL: http://www.iapress.org/index.php/soic/article/view/1202(visited on 03/12/2022).

[25] S M Raju and Ali Mohammad Tarif. “Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis”. In: (2020).Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2006.14473. URL: https://arxiv.org/abs/2006.14473 (visited on 03/12/2022).

[26] Edward Loper and Steven Bird. “NLTK: The Natural Language Toolkit”. In: *arXiv:cs/0205028* (May 2002). arXiv: cs/0205028. URL: http://arxiv.org/abs/cs/0205028(visited on 04/08/2022).

[27] Clayton J. Hutto and Eric Gilbert. “VADER: A Parsi- monious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *ICWSM*. 2014.

[28] Anselm Strauss and Juliet Corbin. “Basics of qualitativeresearch techniques”. In: (1998).