

A Stacking Integrated Learning Framework for Digital Currency Prediction Using Social Media Data

Amone Chanthaphavong¹, Phouthone Vongpasith², Phonesouda Souphamith³,
Tiengthong Phenghachanh⁴, Damduan Kingmaneesengkeo⁵

^{1,2,3,4}Department of Computer Science, Faculty of Natural Sciences, National University of Laos, Vientiane Capital, Lao P. D. R., ⁵Department of Environmental Sciences, Faculty of Environmental Sciences, Vientiane Capital, Lao P. D. R.

Abstract

The development of the internet not only provides convenience for the majority of investors and users to invest in stocks, but also promotes the development of digital currencies through online social media. Although, social media data will also significantly affect the digital currency market, the number of digital currency investors is still very small, due to the lack of information acquisition channels for investors. From a practical point of view, it is valuable to study the relationship between social media data and digital currency and provide a basis for investors' decision-making. Therefore, it is necessary to combine data on multiple mainstream social media platforms to explore its impact on the performance of the digital currency market. Based on the stacking integrated learning framework, this paper studies the impact of social media data on the digital currency market. To obtain the research sample data web crawler technology is used. The development of the influencing factors of data currency prices through the python self-encoding method is conducted. The stacking integrated learning framework used in this paper increases the model's prediction accuracy of sample data by 1.32%.

Key words: Stacking Integrated Learning, Social Media Data, Digital Currency, Data Mining, Web Crawler.

I. INTRODUCTION (HEADING 1)

With the development of technology, people have a deeper understanding of financial knowledge. Their awareness of financial management and investment is gradually enhanced. The number of stock investors is increasing globally [1], but the number of digital currency investors is still very small, due to the lack of information acquisition channels for investors. Before investing in digital currency, investors will comprehensively consider the status of the digital currency market, mainstream media public opinion, and analyst experience analysis. After thinking and judgment, they choose appropriate digital currency for stock investment [2]. The traditional method of information disclosure has the shortcomings of poor timeliness, slow update speed, and complete one-way transmission of information. These shortcomings often make it difficult for investors to understand and obtain the latest status information of the company in real time. In recent years, the internet has developed rapidly all over the world, making information communication channels more diverse, simple and timely [3]. More and more investors understand and become familiar with digital currencies through online media, and further invest in digital currencies [4]. At present, mainstream social media platforms followed by global investors include Facebook, Twitter, Reddit, and CryptoCompare. These social media platforms have a large number of loyal users, including business leaders, investor users, professional analysts, etc. On social media platforms, the person in charge of the company will release daily business information of the company, investor users will browse information,

exchange investment experience, and analyze stock market conditions, resulting in a large amount of valuable social media data [5]. There is no doubt that social media platforms enable investors to obtain more information through more diversified channels, more timely information [6]. From a practical point of view, it is valuable to study the relationship between social media data and digital currency. By quantifying social media data and constructing feature variables, it is possible to predict the trend of the digital currency market and provide a basis for investors' decision-making [7]. The main research purpose of this article is to combine with the feature of the influence of social media data on the rise and fall of digital currency prices, a stacking integrated learning framework is constructed, using mainstream machine learning algorithms[8]; Random Forest, XGBoost, etc. as the first layer of learners, and Support Vector Machine as the second layer of learning device. Through the python package, the experiment compared the accuracy of the stacking integrated learning framework and the previous algorithms in predicting the rise and fall of digital currency prices.

II. RELATED RESEARCH

Blockchain is an optional technology, which is characterized by distributed bookkeeping, it is not based on accounts, and it cannot be tampered. If digital currency focuses on protecting personal privacy, block-chain technology can be used [9]. The issuance of digital currency can reduce the high cost of traditional paper currency issuance and circulation, it also can reduce money laundering, tax evasion and other illegal crimes. Moreover, it can improve the central bank's currency supply and the convenience and transparency of economic transaction activities[10],[2].Bitcoin completes the whole issuance process by distributing newly generated coins as block rewards to miners [11]. There are strict established rules for issuance, and no one has the right to modify these rules and make additional issuance outside of the rules [12], [13]. The risk is basically controllable, and technical issues that are not the most important issue currently facing digital currencies [14]. The security core of most digital currencies is based on a specific hash algorithm. If the algorithm is cracked, an attacker can derive the same information digest from two different sets of information [15], [16]. In addition to large off-market transactions, trading platforms are where most digital currency enthusiasts buy and sell digital [17]. Due to the proximity to the money and the relatively large amount of funds in the account, the trading platform is the most vulnerable link in the entire digital currency system. Although many countries have recognized the status of Bitcoin and assigned it different status, in general, the relevant laws are not sound enough [18]. Social media data mining has also become one of the most active research directions in the field of data mining and knowledge discovery. The current status and trends of online public opinion in the social media era, and pointed out that online public opinion can affect frequency trading behavior [19], [20] To analyzing the big data in finance and to predict the possible trading behavior of the participant during the rest of the trading hours are performed the same behavior [22]. Market sentiment has an important impact on financial markets [21]. Through the previous literature, it is found that the risk-related research of digital currency is relatively mature. Price fluctuations, digital currency itself has technical risks, the use of digital currency in wallets, hash algorithms being cracked, attack issues, trading platform risks, and digital currency policy risks are the main risks of digital currency proposed by current research.

III. STACKING INTEGRATED LEARNING FRAMEWORK

The stacking method first uses different base classifiers to learn the original data, and then uses the output value of each base learner as the input of the two-layer learner, and the two-layer learner obtains the final model output value. Based on the stacking integrated learning framework, the first layer of the model includes logistic regression algorithms such as, CART (Classification and Regression Trees), RF (Random

Forest), GBDT (Gradient Boosting Decision Tree) and XGBoost, and the second layer uses SVM (Support Vector Machine).

IV. DATA ACQUISITION

In terms of data acquisition, this article first adopts the web crawler method and uses self-encoding to acquire social media platform data. A total of 929,495 pieces of data currency-related information were collected on Facebook, Twitter, Reddit, and CryptoCompare, a total of 3,717,980 pieces of data currency-related information. Secondly, obtain the relevant data of the digital currency market from the coinmarketcap website through a web crawler method. The web crawler code is used to crawl the basic information of the top 800 digital currencies from the coinmarketcap website, and store them in the MySQL database. The basic information of digital currency mainly includes four columns: ID, Name, Symbol, and Slug, as shown in Table 1.

Table 1 Basic Information of Digital Currency

<i>ID</i>	<i>Name</i>	<i>Symbol</i>	<i>Slug</i>
1	Bitcoin	BTC	Bitcoin
2	Ethereum	ETH	Ethereum
3	Tether	USDT	Tether
4	Litecoin	LTC	Litecoin
5	Namecoin	NMC	Namecoin

After obtaining the basic information of the digital currency, the transaction data of these digital currencies on the coinmarketcap website during the time interval from January 1, 2019 to June 30, 2020 is obtained using the web crawler technology.

Table 2 An Example of Digital Currency (Bitcoin) Market Data

<i>Date</i>	<i>Open</i>	<i>Max</i>	<i>Min</i>	<i>Close</i>	<i>Cnt</i>	<i>Value</i>
Jun 30, 2020	9,185.58	9,217.84	9,084.84	9,137.99	15,735,797,744	168,315,606,321
Jun 29, 2020	9,140.03	9,237.57	9,041.88	9,190.85	16,460,547,078	169,280,659,246
Jun 28, 2020	9,048.46	9,197.55	8,975.53	9,143.58	14,560,870,760	168,401,806,137

V. COMPARISON AND RESULT ANALYSIS

This article quantifies the social media data, which can be specifically divided into two aspects: user attention and account activity. The results of the correlation test between the social media feature variables and digital currency trading volume are shown in Table 3.

Table 3 Correlation test between Feature variables and digital currency trading volume

<i>No</i>	<i>Symbol</i>	<i>Abbreviation</i>	<i>Corr</i>	<i>Sig</i>
1	FUA	User attention on Facebook platform digital currency related topics	0.439***	0.000
2	FUL	User likes on topics related to digital currency on the Facebook platform	0.566***	0.000
3	TAF	Number of followers of Twitter platform digital currency official	0.557***	0.000

<i>No</i>	<i>Symbol</i>	<i>Abbreviation</i>	<i>Corr</i>	<i>Sig</i>
		account		
4	TAP	Twitter platform digital currency official account information released	0.519***	0.00 0
5	TUP	Clicks on Twitter platform digital currency official accounts	0.511***	0.00 0
6	TUL	Twitter platform digital currency official account release information user likes	0.251***	0.00 0
7	TLT	Mentions of topics related to digital currency on Twitter	0.188***	0.00 0

Note: *** means $p < 0.001$, ** means $p < 0.01$, * means $p < 0.05$, Sig means Significance, Corr means Correlation

It can also be known from the Table 3 that the correlation between the quantified feature variables and digital currency trading volume of Facebook and Twitter is greater than of the Reddit and CryptoCompare platforms. The four categories of social media data have obvious contributions to the digital currency market prediction. The importance of the feature variables related to the Facebook and Twitter platform data is not much different, and both are greater than the importance of the feature variables related to the Reddit platform. At the same time, the features related to the variables in CryptoCompare platform are the least important.

Table 4 is the precision rate, recall rate and F1 fluctuation prediction of the stacking integrated learning framework. The precision and recall indicators in this article both exceed 0.72, and the F1 value is 0.745. It can be seen that for the sample data in this article, the accuracy of the model prediction is about 74.5%, and the stacking integrated learning framework has a high prediction accuracy, which can be applied in practice. Therefore, based on the stacking integrated learning framework using social media data to predict digital currency market indicators, the prediction results have certain practical significance.

Table 4 Calculation Results of Various Indicators

<i>Model</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Stacking Integrated Learning Framework	0.762	0.729	0.745

In order to further evaluate the prediction effect of stacking integrated learning framework, mainstream data mining algorithms is compared. The results are shown in Table 5. It can be seen from the table that, except for the low accuracy of the LR algorithm, the accuracy of the other algorithms for predicting the stock price rise and fall is more than 70%. The XGB algorithm has the highest accuracy among the commonly used algorithms, reaching 75.07%. The Stacking integrated learning framework has the highest accuracy in predicting stock price fluctuations, reaching 76.39%, which is 1.32% higher than the accuracy of the XGB algorithm. The experimental results show that among the experimental data used in this paper to predict digital currency price data using social media data, the model prediction effect of the stacking integrated learning framework is the best.

Table 5 A comparison of prediction accuracy

<i>Algorithms</i>	<i>LR</i>	<i>CART</i>	<i>RF</i>	<i>GBDT</i>	<i>XGB</i>	<i>Stacking</i>
Accuracy	0.6124	0.7319	0.7131	0.7374	0.7507	0.7639

VI. CONCLUSION

Based on the stacking integrated learning framework, this article is used social media data to predict the price of digital currencies. The first layer uses CART, RF, GBDT, and XGB algorithms with excellent prediction effects, and effectively uses the prediction advantages of these algorithms directly on complex data sets. The second layer uses SVM The algorithm effectively utilizes the predictive advantages on simple data sets. The constructed two-layer model fusion framework effectively improves the accuracy of model prediction. The feature variables related to digital currency market data contribute to the prediction effect of the model far greater than those related to social media data. The importance of the feature variables related to the Facebook and Twitter platform data is greater than the importance of the feature variables related to the Reddit platform data, while the features related to the Crypto compare platform data variables are the least important. This article failed to dig deeper into the emotions contained in the social media data, and the emotions contained in the texts on the platforms can effectively reflect users' sentiments towards the digital currency market. Therefore, future scholars can introduce user sentiment contained in such social media texts into relevant research.

VII. ACKNOWLEDGMENT

The authors would like to express their gratitude to the anonymous referees for their valuable comments and suggestions.

REFERENCES

- [1] G. Fonseca, "An analysis of the legal impact of central bank digital currency on the European payments landscape," *Journal of Payments Strategy & Systems*, vol. 13, no. 4, 2019.
- [2] J. Groß, M. Klein, and P. Sandner, "Digitale Zentralbankwährungen: Chancen, Risiken und Blockchain-Technologie [Central Bank Digital Currencies: Benefits, Risks and the Role of Blockchain Technology]," *Wirtschaftsdienst*, vol. 100, no. 2020.
- [3] C. Guo and H. Li, "Study on the big data capacity building of internet public opinion governance in the new media environment," *Information Theory and Practice*, vol. 41, no. 12, pp. 46-54, 2018.
- [4] K. Wang and W. Yang, "An online social media location inference method under big data environment," *Journal of Software*, vol. 26, no. 11, pp. 2951-2963, 2015.
- [5] B. Li and M. Zhao, "Internet public opinion in the social media era: Ecological changes and public opinion research status and trends," *Journalists*, vol. 1, pp. 36-41, 2014.
- [6] J. Zhao and Z. Xu, "Big data era: The status quo and trends of media development," *Media*, vol. 21, no. 9, pp. 47-50, 2013.
- [7] L. Weng and S. Xu, "Portfolio trading system of digital currencies: A deep reinforcement learning with multidimensional attention gating mechanism," *Neurocomputing*, vol. 402, pp. 171-182, 2020.
- [8] O. Hutsady and P. Vongpasith, "Lao Telecom Customer Churn Prediction: A Case Study in Borikhamxay Brunch," *International Journal of Scientific Development and Research*, vol. 9, no. 8, pp. 296-400, 2024.
- [9] L. Jia, "The theory, practice and influence of Bitcoin," *Research in International Finance*, vol. 12, pp. 14-25, 2013.
- [10] G. Fonseca, "An analysis of the legal impact of central bank digital currency on the European payments landscape," *Journal of Payments Strategy & Systems*, 2020.
- [11] Y. Fan, "Theoretical basis and structure selection of China's legal digital currency," *China Finance*, vol. 17, pp. 10-12, 2016.

- [12] R. Grinberg, "Bitcoin: An innovative alternative digital currency," *Hastings Sci. & Tech. LJ*, vol. 4, pp. 159, 2012.
- [13] S. Gilbert and L. Huang, "Digital currency risk," *International Journal of Economics and Finance*, vol. 10, no. 2, pp. 108, 2018.
- [14] N. Kaplanov, "Nerdy money: Bitcoin, the private digital currency, and the case against its regulation," *Loy. Consumer L. Rev.*, vol. 25, pp. 111, 2012.
- [15] A. M. Zacarias and B. A., "Digital exchange and mobile wallet for digital currency: U.S. Patent 8,768,834," 2014.
- [16] A. Lipton and T. H., "Digital trade coin: towards a more stable digital currency," *Royal Society Open Science*, vol. 5, no. 7, p. 180155, 2018.
- [17] R. Auer and B. R., "The technology of retail central bank digital currency," *BIS Quarterly Review*, 2020.
- [18] T. Tan and Y. S., "On the transformation of social media relations," *Modern Communication, Journal of Communication University of China*, vol. 35, no. 11, pp. 108-113, 2013.
- [19] Y. Wen, "Research on social media short text classification method," M.S. thesis, University of Electronic Science and Technology of China, 2018.
- [20] J. Shi and L. X., "Simulation of tourists' spatiotemporal behavior and result validation with social media data," *Transportation Planning and Technology*, vol. 43, no. 7, 2020.
- [21] R. Samuels and E. J., "Deepening the divide: Crises disproportionately silence vulnerable populations on social media," *Journal of Management in Engineering*, vol. 36, no. 6, 2020.
- [22] C. Sharma and W. S., "Mining social media data to investigate patient perceptions regarding DMARD pharmacotherapy for rheumatoid arthritis," *Annals of the Rheumatic Diseases*, 2020.