LSTM Based Sentiment Analysis for Cryptocurrency Prediction

Xin Huang[1], Wenbin Zhang[1], Xuejiao Tang[2], Mingli Zhang[3], Jayachander Surbiryala[4], Vasileios Iosifidis[2], Zhen Liu[5] and Ji Zhang[6]

[1] University of Maryland, Baltimore County, USA
[2] Leibniz University Hannover, Germany
[3] McGill University, Canada [4]University of Stavanger, Norway
[5]Guangdong Pharmaceutical University, China
[6]University of Southern Queensland, Australia
[1]{xinh1,wenbinzhang}@umbc.edu,
[2]{xuejiao.tang,iosifidis}@stud.uni-hannover.de,
[3]mingli.zhang@mcgill.ca, [4]jayachander.surbiryala@uis.no,
[5]liu.zhen@gdpu.edu.cn, [6]ji.zhang@usq.edu.au

Abstract. Recent studies in big data analytics and natural language processing develop automatic techniques in analyzing sentiment in the social media information. In addition, the growing user base of social media and the high volume of posts also provide valuable sentiment information to predict the price fluctuation of the cryptocurrency. This research is directed to predicting the volatile price movement of cryptocurrency by analyzing the sentiment in social media and finding the correlation between them. While previous work has been developed to analyze sentiment in English social media posts, we propose a method to identify the sentiment of the Chinese social media posts from the most popular Chinese social media platform Sina-Weibo. We develop the pipeline to capture Weibo posts, describe the creation of the crypto-specific sentiment dictionary, and propose a long short-term memory (LSTM) based recurrent neural network along with the historical cryptocurrency price movement to predict the price trend for future time frames. The conducted experiments demonstrate the proposed approach outperforms the state of the art auto regressive based model by 18.5% in precision and 15.4% in recall.

1 Introduction

Since the birth of Bitcoin, there has been an enormous rise and interest in the cryptocurrency, a decentralized digital asset developed by the blockchain technology. This digital currency draws a lot of attention due to its volatility which provides the opportunity for digital trading with high return. The total market capitalization of cryptocurrencies has increased from 1 billion dollars to 400 billion dollars in the past decade, with the number still increasing.

On the other hand, the emergence of social media such as Twitter, Reddit and Facebook also makes the latest news and social media posts about financial markets widely accessible. Investors have therefore been utilizing such a variety of digital resources to make trading decisions. Previous studies discovered evidence

of such correlation between stock price movement and social media [1]. Sentiment on cryptocurrency social media content with negative emotions, e.g., fear and sadness, neutral emotions, e.g., calm and not sure, or positive emotions, e.g., trust and happiness, can be used to predict cryptocurrency price fluctuations and further to assist the investment decision making. This paper focuses on this trending theme, proposing a recurrent neural network with long short-term memory (LSTM) by utilizing the sentiment analysis of social media to predict the real time price movement of the digital currency.

2   Related Work

Over the past decades, a variety of machine learning techniques have been developed to predict the price movement for the stock market using social media, such as opinion analysis of twitter feeds [3,13,12,10,8,5]. [1] used neural networks and daily Twitter feeds as extra predictors to forecast the daily up and down changes in the closing values of the Dow Jones Industrial Average. Moreover, [9] found the Long Short-Term Memory (LSTM) combined with a Twitter sentiment analysis outperforms other machine learning models such as Support Vector Machine in predicting the stock price.

Recent studies have also successfully applied sentiment analysis in various applications, such as predicting the movie revenues [7], analysing sentiment towards US presidential candidates in 2012 [11]. In the English sentiment analysis, Valence Aware Dictionary and sEntiment Reasoner (VADER) [6] is used to classify sentiment in tweets. In this paper, we designed a crypto sentiment dictionary that is customized to cryptocurrency and Chinese Weibo posts. We use LSTM [4] as the neural network learning layer and combine it with the sentiment analysis method to develop a crypto sentiment analyzer that can predict the price movement of cryptocurrency.

3   Methodology

Figure 1 shows the end to end architecture of the LSTM based sentiment deep learning model in predicting the real time price fluctuations. We crawled user posts from China's most popular social media platforms Sina-Weibo, and created a crypto-specific sentiment dictionary with domain-expert knowledge, and then the LSTM recurrent neural network was used to model the sentiment information and make real time prediction for the price trend.

3.1   Data Collection

Chinese investors exchange crypto information via news articles and social media platforms, especially using Sina-Weibo, Wechat and QQ groups. We collect a large-scale Weibo corpus from crawling Chinese microblogs on Sina-Weibo with the cryptocurrency keyword, in particular, Bitcoin, ETH or XPR. The number of crawled cryptocurrency tweets from Weibo is 24,000, as well as 70,000 comments to them, from the most recent 8 days.
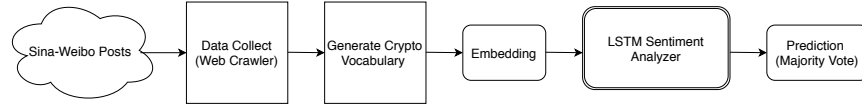
Fig. 1. Architecture of LSTM based cryptocurrency sentiment analysis and price movement prediction. Crypto sentiment dictionary is created to generate the crypto word embedding, LSTM is to learn the sentiment information, and the majority voting on the output of LSTM sentiment analyzer is used to predict the price going up or down.

3.2 Crypto Sentiment Dictionary

The general sentiment dictionary created by natural language processing (NLP) is not applicable in the crypto domain. We introduce a novel way to build a crypto specific sentiment dictionary that can capture the unique characteristics of the crypto social communities. Table 1 shows an example dictionary that is particular to the Chinese crypto words in Sina-Weibo. The first step of generating a crypto sentiment dictionary is to create the vocabulary of the crypto words. We manually label the crawled Weibo posts with ranking, and use seed sentiment words selected by crypto domain experts to do bootstrapping, which adds high frequency new words in highly positive/negative weibo posts into the crypto corpus.

After the crypto corpus is generated, we then create an index mapping dictionary in such a way that the frequently occurring crypto words are assigned lower indexes, similar to the traditional natural language processing. Finally we generate a crypto word encoding for each individual post and use that encoding vector as the training data for the RNN model in sentiment analysis.

Table 1. Crypto specific sentiment dictionary building for Chinese words.

Chinese	Informal Translation	Implied Sentiment
韭菜	Bag Holder	Investors have profit loss
新高	New High	Investors are excited
下车	Abandon Ship	Investors rush to sell off

3.3 LSTM based Sentiment Analyzer

We develop a long short-term memory network (LSTM) based sentiment analyzer for crypto social media posts. LSTM enables the network to learn long-term relation, by utilizing forget and remember gates that allow the cell to decide which information to block or transmit based on its strength and importance.

The social media post is first tokenized according to crypto word vocabulary and fed into the embedding layer, which converts the word token into the crypto word embedding. The LSTM based recurrent network is trained by taking the sequence of the embedding feature vector. A fully connected layer is used to transform the output of the LSTM later and activated with sigmoid to output the prediction. The labels of the posts used in training were manually labeled and encoded with positive (1), neutral (0) and negative (-1).

4   Evaluation

We used the most recent 7 days' Sina-Weibo posts from top 100 crypto investors accounts as training data and the next 1 day's posts as testing. We use Precision and Recall to measure the performance of our LSTM sentiment predictor. Precision measures the model's ability to return only relevant instances and recall measures the model's ability to classify all relevant instances.

We compare our method with the time series auto regression (AR) approach [2] to evaluate the performance. As Table 2 shows, our approach outperforms the AR approach by 18.5% in precision and 15.4% in recall, exemplifies the effectiveness of the LSTM in analyzing the sentiment of social media content.

Table 2. Precision and recall on evaluating LSTM sentiment analyzer and AR.

Method	**Precision**	**Recall**
Auto Regression	73.4%	80.2%
LSTM Sentiment Analyzer	87.0%	92.5%

5   Conclusion

In this paper we introduce a crypto sentiment analyzer by utilizing the recurrent neural network to model the social media sentiment. The model is developed using LSTM and achieves higher precision and recall than the traditional auto regressive approach. The current sentiment analyzer can be used to predict the price fluctuation of the cryptocurrency and integrated to an autonomous trading system to assist the buying or selling of digital assets.

References

1. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**, 1–8 (2011)
2. Box, G., Jenkins, G., Reinsel, G., Ljung, G.: Time series analysis : forecasting and control. In: ISBN 1-118-67502-9. OCLC 915507780 (2016)
3. Chen, R., Lazer, M.: Sentiment analysis of twitter feeds for the prediction of stock market movement. Stanford Computer Science **229** (2011)
4. Hochreiter, S., Schmidhuber, J.: Lstm can solve hard long time lag problems. In: Proceedings of the 9th International Conference on Neural Information Processing Systems. pp. 473–479 (1996)
5. Huang, X., Tang, X., et al.: Research on knowledge based expert system for medical diagnosis (2021)
6. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media. pp. 216–255 (2014)
7. Joshi, M., Das, D., Gimpel, K., Smith, ., a. N.: Movie reviews and revenues: An experiment in text regression. human language technologies. In: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 293–296 (2010)
8. Liu, Z., Wang, R., Japkowicz, N., Tang, D., Zhang, W., Zhao, J.: Research on unsupervised feature learning for android malware detection based on restricted boltzmann machines. Future Generation Computer Systems **120**, 91–108 (2021)

9. Pimprikar, R., Ramachadran, S., Senthilkumar, K.: Use of machine learning algorithms and twitter sentiment analysis for stock market prediction. International Journal of Pure and Applied Mathematics pp. 521–526 (2017)
10. Tang, X., Huang, X., Zhang, W., Child, T.B., Hu, Q., Liu, Z., Zhang, J.: Cognitive visual commonsense reasoning using dynamic working memory. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 81–93. Springer (2021)
11. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: Hand gesture recognition using fourier descriptors. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. pp. 115–120 (2012)
12. Zhang, W., Zhang, L., Pfoser, D., Zhao, L.: Disentangled dynamic graph deep generation. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). pp. 738–746. SIAM (2021)
13. Zhang, W., Zhao, L.: Online decision trees with fairness. arXiv preprint arXiv:2010.08146 (2020)