Research on Tweet Sentiment Analysis Based on VADER in the Field of Cryptocurrency

Ji Miao

College of Information Science and Technology, Hangzhou Normal University, Hangzhou, Zhejiang Province 310036, China

E-mail: 1060150683@qq.com, Website: www.hznu.edu.cn

Abstract. This research aims to develop a tool to assess the impact of social media on the market by collecting and analyzing cryptocurrency-related tweets on Twitter. Python and relevant libraries are utilized for scraping tweets and cryptocurrency trading data, followed by data preprocessing. The VADER model is then employed for sentiment analysis of the tweets, extracting sentiment polarity and intensity. Considering the influence of tweets, an overall sentiment score is calculated.

Keywords: Cryptocurrency, Opinion Risk Monitoring, Sentiment Analysis, VADER, Cross-correlation Analysis

1 Introduction

Blockchain is a decentralized and anonymous infrastructure that has emerged alongside cryptocurrencies. Its characteristics, such as the trust mechanism and incentive mechanism, have expanded its applications from digital currencies to various fields, including e-commerce and the internet.

Social media platforms like Twitter and Reddit have become vital channels for the rapid dissemination and discussion of cryptocurrency-related information. Influential figures in the digital currency community, with massive followings, can significantly impact cryptocurrency prices with just a single tweet or post.

This study aims to develop tools to assess the impact of social opinion on market trends by collecting cryptocurrency-related tweets on Twitter. The sentiment tendency and user influence of these tweets will be analyzed using sentiment calculation and VADER model algorithms. By establishing a quantitative relationship with actual cryptocurrency price changes, we will focus on analyzing the emotional polarity of tweets, the number of retweets, and likes.

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_28

2 Research methodology

2.1 Data Pre-processing

2.1.1 Getting Tweets

First, import the tweepy library, which facilitates access to the Twitter API. Set up authentication credentials for the Twitter API by obtaining your user key, user secret, access token, and access token secret from your Twitter developer account.

Next, pass the authentication information to tweepy to create an API connection object. Define your search criteria, which involves using keywords, hashtags, and handles related to Bitcoin. For example, "#bitcoin," "#btc," "bitcoin," "btc."Utilize tweepy's Cursor object to search the API using your defined criteria. Iterate through the results and store the tweets. It's important to note that Twitter's API limits calls to 180 requests per 15 minutes. When this limit is reached, the script should enter a sleep state and handle exceptions. After the sleep period, continue crawling.Finally, for each tweet, extract the following information: tweet ID, tweet content, username, user's follower count, retweet count, like count, and creation date.

2.1.2 Obtaining BTC Data

First, register a free API key, pass it as a parameter in the API request, and define the API endpoint URL for Bitcoin data. Create a dictionary and send a GET request.

Next, parse the JSON response using the json module. Access specific fields in the JSON data, including the timestamp: the time the data was issued, closing price: the price at the end of the time range (e.g., per minute, per hour, or per day, depending on the target endpoint), highest price: the highest price reached within the time range, lowest price: the lowest price reached within the time range, and opening price: the price at the start of the time range.Finally, to enable correlation analysis with tweets, ensure that the time range of the BTC transaction data obtained here overlaps with the time range of the tweets. You can pass additional parameters to achieve this.

2.1.3 Tweet Preprocessing

First, unify the letter case of all tweet data and remove emojis to normalize the text. Employ methods like Regex to eliminate URLs, Twitter handles, #hashtags, punctuation, as well as stop words such as "a," "the," and "and," which do not contribute to semantic value.

Next, after preprocessing the collected Twitter data, we perform sentiment analysis on the plain text content of each tweet. Utilize natural language processing techniques to assign a sentiment score to each tweet, indicating the overall positive, negative, or neutral sentiment expressed in the tweet text.

2.1.4 BTC Transaction Data Preprocessing

First, check and remove rows in key fields (such as transaction ID, sender, receiver, amount, etc.) that have missing or blank values. Next, conduct additional

validation checks on addresses and transaction types to ensure the amounts/balances are reasonable. Finally, eliminate records where key fields are duplicated.

2.2 Models and Algorithms

2.2.1 Sentiment analysis

After data preprocessing, we utilized the VADER (Valence Aware Dictionary and sEntiment Reasoner) model to perform sentiment analysis on each tweet. VADER is a dictionary-based and rule-combining model that captures the polarity and intensity of sentiments in text. We chose to use VADER for the following reasons:

VADER has been optimized specifically for social media text and can handle characteristics of online language, such as abbreviations and emojis.

VADER categorizes text into positive, neutral, and negative polarities and provides sentiment intensity scores, making it more comprehensive.

VADER combines both dictionary and rule-based approaches, resulting in better performance compared to purely dictionary-based methods.

The specific steps are as follows:

Load a pre-trained VADER model from GitHub.

For each tweet, use the model to generate positive, negative, and neutral scores.

Calculate a composite score as the difference between the positive score and the negative score, within a range of -1 to 1.

Determine the sentiment polarity and strength of each tweet based on the composite score, considering the influence of the tweet.

Weighted average the sentiment scores of all tweets to obtain an overall sentiment score. This allows us to analyze the overall sentiment trends regarding cryptocurrency on Twitter.

Next, we conducted correlation analysis. To quantify the correlation between tweet sentiment and cryptocurrency price changes, we employed cross-correlation analysis. The main advantages of this approach include:

Detecting lagged correlations, meaning changes in tweet sentiment precede price changes.

Handling both linear and non-linear relationships.

Providing correlation coefficients and lag times.

The specific steps for this analysis are as follows:

Calculate cross-correlation coefficients at different lags between the tweet sentiment time series and the cryptocurrency price time series.

Identify the lag time (τ) with the highest absolute cross-correlation coefficient (R).

Perform hypothesis testing to determine if the correlation is statistically significant.

Analyze the results to understand whether changes in tweet sentiment precede price changes and influence prices.

This methodology allows us to explore and quantify the relationship between sentiment expressed on Twitter and cryptocurrency price fluctuations.

2.2.2 Vector Autoregression

Apply the vector autoregressive (VAR) model to estimate dynamic relationships between variables. Construct equations using lagged values of endogenous variables. Build multiple equations to describe interactions between opinion risk, tweets, and cryptocurrency trading data.

3 Results

3.1 BTC

We analyzed Bitcoin's correlation with tweets and cryptocurrencies based on their derivatives.(in Figure 1) Peaks in cryptocurrency derivatives indicate significant increases, while flat derivatives (That is, where the blue line is 0 in Figure 2) suggest stabilization.

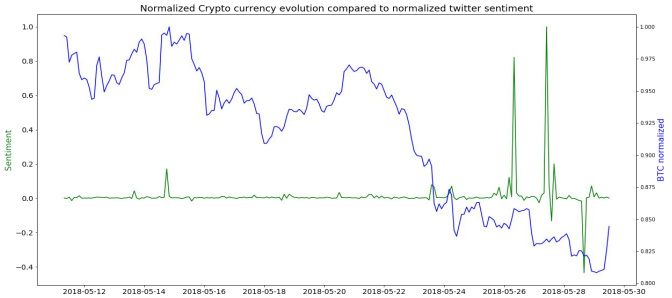

Fig. 1. Tweets and cryptocurrencies

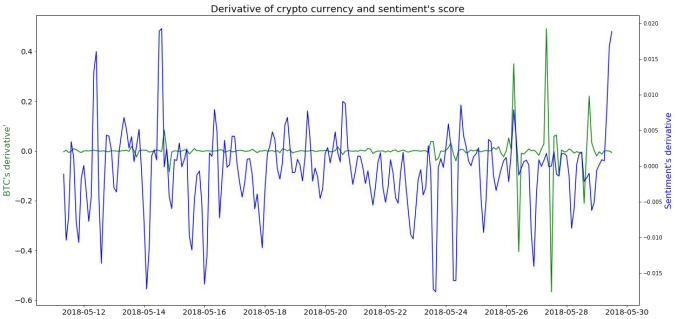

Fig. 2. Tweets and cryptocurrencies - derivatives

It is difficult to see the correlation in Figure 2, where we find that derivatives of cryptocurrencies have a large magnitude compared to derivatives of sentiment. Finally there seem to be very few peak matches, but these are outliers and we cannot define rules based on these impressions. That's why we calculate correlations using pandas' dataframe built-in method corr, which allows us to correlate in Pearson, Kendall and Spearman correlations. When we work with time series and

cryptocurrencies, we should add a lag parameter to move one of the series to the left or right in order to expect a higher correlation. This is called a cross-correlation. In Figure 3, we change the lag between -20 and +20, and since the data is grouped by 2 hours, this means that the sentiment series switches between -40 and +40 hours.

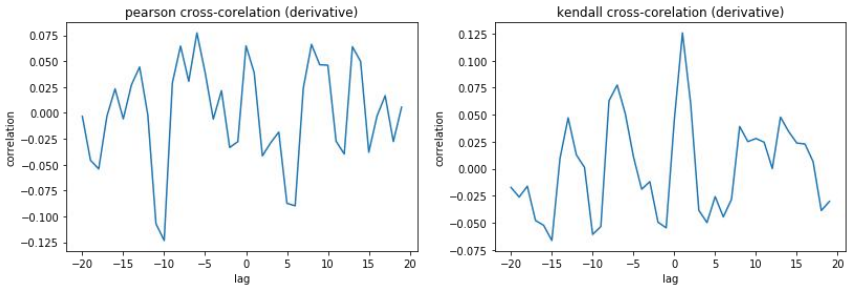

Fig. 3. Degree of interrelationship

A correlation of +1 or -1 means that both sentiment and cryptocurrency derivatives match, so we can use this as a predictive mechanism. A correlation of 0 means absolutely no match.Kendall and Spearman look very similar, but Spearman gets the best score: 0.2 with a lag of 1. However, the results are meaningless and the correlation is too low to infer any predictions. The Spearman correlation between two variables is equal to the Pearson correlation between the ranked values of the two variables; the Pearson correlation assesses a linear relationship, while the Spearman correlation assesses a monotonic relationship (linear or not). Therefore, we believe that Spearman is more suitable for time series data where the relationship is nonlinear.

3.2 Zilliqa

We tested the same at Zilliqa. On this chart(Figure 4), positive sentiment seems to lead to currency depreciation. An interesting rule for making money is the opposite of what people think.

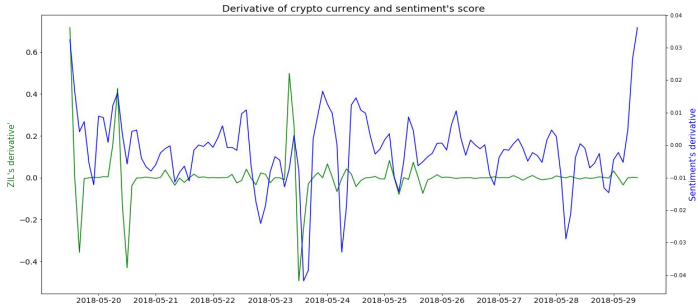

Fig. 4. Zilliqa and tweets

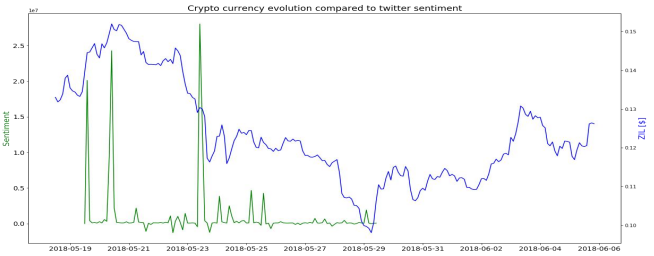

Fig. 5. Zilliqa and tweets - derivatives

A stronger correlation between this currency and tweets seems to be seen on Figure 5. In fact, we obtain a score of 0.3 with a lag of 0, see Figure 6.

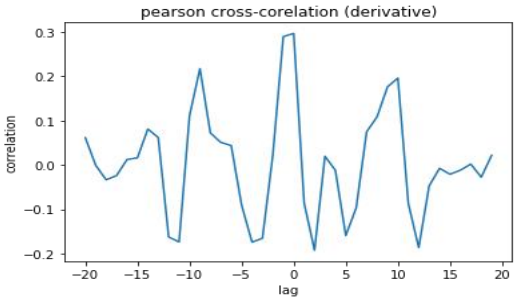

Fig. 6. Degree of interrelationship

3.3 Comparing BTC and Zilliqa

If we compare 2 charts of different cryptocurrencies. We can observe that most of the time, for each cryptocurrency, when the Twitter sentiment goes up significantly, the cryptocurrency goes down. This could be due to the fact that people see the hype and so they decide to sell. In a few cases, we observe that when Twitter has a negative sentiment, the price goes up. the LSTM may be able to use these big spikes to decide when to sell.

3.4 Vector autoregression

Table 1. ADF test

variant	t	P	threshold value		
			1%	5%	10%
emotional score	-3.289	0.015**	-3.449	-2.87	-2.571
Transaction data	-8.086	0.000***	-3.448	-2.869	-2.571
public opinion risk	-3.226	0.019**	-3.449	-2.87	-2.571

Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.

Based on the variable sentiment score, the significance p-value is 0.015**, presenting significance at the level, rejecting the original hypothesis that the series is a smooth time series.(in Table 1) Based on the variable Transaction Data, the significance p-value is 0.000***, presenting significance at the level, rejecting the original hypothesis, the series is a smooth time series. Based on the variable Public Opinion Risk, the significance p-value is 0.019**, presenting significance at the level, rejecting the original hypothesis, the series is a smooth time series.

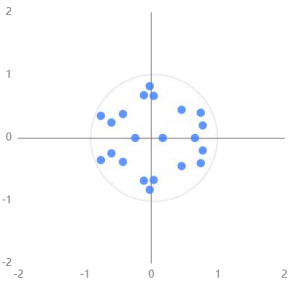

Fig. 7. VAR model stability test

Figure 7 illustrates the AR root plot in the VAR model. If all the points are located within the unit circle, thus the VAR system can be judged to be stable and the model can be further impulse response analysis and variance decomposition.

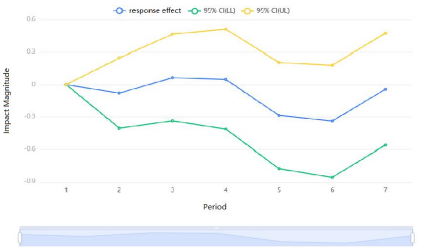

Fig. 8. Impulse response analysis graph

In the VAR model, the impact of a positive tweet on the risk of public opinion changes from a negative response to a positive response in Figure 8, and there is an obvious convergence only in the 7th period, indicating that the tweets have a more stable and lasting impact on the analysis of public opinion. The negative and then positive response is also in line with the daily situation that when a major figure in the cryptocurrency circle sends out a positive tweet, outsiders' first reaction is to wait for the overall situation to change before deciding whether or not to enter the market, and insiders will be a little bit panicked, and only after a couple of issues have passed will they gradually tend to be positive and reach a convergence.

4 Summarize

This study examines the relationship between tweet sentiment and cryptocurrency price changes, focusing on Bitcoin. Twitter data with relevant keywords were collected using Python. Historical cryptocurrency trading data was also gathered. After preprocessing, the VADER model gauged tweet sentiment. Considering user influence, tweets were weighted to calculate an overall sentiment score time series. Cross-correlation analysis was used to assess the lagged correlation between tweet sentiment and cryptocurrency prices. While tweet sentiment often anticipated price changes, the correlation wasn't strong enough for predictive models. Future work could involve deep learning models for a more comprehensive analysis. In essence, the study confirms social media's potential impact on prices, laying a foundation for an opinion risk monitoring system.

Bibliography:

1. Nakamoto, Satoshi. Bitcoin A Peer-to-Peer Electronic Cash System. - References - Scientific Research Publishing [EB/OL]. [2023-07-17]. https://www.scirp.org/(S(oyulxb452alnt1aej1nfow45))/reference/ReferencesPapers.aspx?ReferenceID=1522950.
2. Buterin V. A NEXT GENERATION SMART CONTRACT & DECENTRALIZED APPLICATION PLATFORM[C]. 2015.
3. Study on the Path and Countermeasures of Blockchain Technology to Promote High-Quality Development of Digital Economy in Guangzhou [EB/OL]. [2023-07-17]. https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/journal/paperinformation.aspx?paperid=121332.
4. Chen Y, Bellavitis C. Blockchain disruption and decentralized finance: the rise of decentralized business models[J]. Journal of Business Venturing Insights, 2020, 13: e00151.
5. Yan Xiaofang. Dilemmas and Countermeasures Suggestions for Online Public Opinion Monitoring in the Context of Mobile Internet[J]. Media, 2021(8): 74-76.
6. Zhou M (Jamie), Lei L (Gillian), Wang J, et al. Social Media Adoption and Corporate Disclosure[J]. Journal of Information Systems, 2015, 29(2): 23-50.
7. Guidi B. When Blockchain meets Online Social Networks[J]. Pervasive and Mobile Computing, 2020, 62: 101131.
8. Ripley D M. Systematic Elements in the Linkage of National Stock Market Indices[J]. The Review of Economics and Statistics, 1973, 55(3): 356-361.
9. Luo Peng, Chen Yiguo, Xu Chuanhua. Baidu Search, Risk Perception and Financial Risk Prediction - A Perspective Based on Behavioral Finance[J]. Financial Forum, 2018, 23(1): 39-51.