Expert Systems With Applications 219 (2023) 119600

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-source data driven cryptocurrency price movement prediction and portfolio optimization

Zhongbao Zhou [a,b], Zhengyang Song [a,*], Helu Xiao [c], Tiantian Ren [a]

[a] School of Business Administration, Hunan University, Changsha 410082, China
[b] Institute of Data Science and Decision Optimization, Hunan University, Changsha 410082, China
[c] School of Business, Hunan Normal University, Changsha 410081, China

ARTICLE INFO

Keywords:
Cryptocurrency
Multi-source data
Portfolio optimization
Support Vector Machine
Out-of-sample test

ABSTRACT

The existing cryptocurrency portfolio studies have relied heavily on historical asset returns and ignored the importance of the prediction information of asset returns, which leads to poor out-of-sample performance of the resulting portfolio strategies. To this end, we first crawl the tweets related to cryptocurrencies on Twitter, analyze tweets' sentiment, and construct sentiment indicators. Second, we use the historical trading data, daily Google Trends, and sentiment indicators to forecast the movement of cryptocurrency prices using Support Vector Machine (SVM). Third, we propose a portfolio optimization model by considering both the forecasting information and the global minimum variance model, and then derive the corresponding portfolio strategy. Finally, we compare the out-of-sample performance of the proposed strategy with classic portfolio strategies and the Cryptocurrency Index. The empirical results show that: on the one hand, the proposed multi-source data can effectively help forecast the cryptocurrency price movements; on the other hand, the proposed portfolio strategy outperforms traditional portfolio strategies regarding the out-of-sample Sharpe ratio, Sortino ratio, and certainty equivalent return, this proves that the proposed strategy can sufficiently combine information between history and future. More importantly, the above conclusions are well verified in the robustness test.

1. Introduction

Over the past few years, cryptocurrencies have grown as fast as a meteor, reaching record highs and a total market capitalization of nearly $2 trillion as early 2022. Furthermore, many countries such as Germany, Japan, the United States, and South Korea have authorized the legal payment status of cryptocurrencies. Bitcoin, the first decentralized cryptocurrency introduced by Nakamoto (2008), accounts for about 40 % of the overall cryptocurrency market. With the success of Bitcoin, more and more cryptocurrencies have been created, and nowadays, emerging cryptocurrencies such as Ethereum, Ripple, Cardano, and Dogecoin are becoming increasingly popular among investors. There is empirical evidence that Bitcoin and other cryptocurrencies should be treated not as a currency but as speculative assets (see, e.g., Glaser et al., 2014; Baek and Elbeck, 2015; Makarov and Schoar, 2020). In addition, many investors and scholars have considered cryptocurrencies as a new alternative investment and are trying to obtain lower risk and higher returns through the diversity of cryptocurrency portfolios (see Guesmi et al., 2019; Liu, 2019; Bouri et al., 2020).

To the best of our knowledge, many scholars have considered cryptocurrencies in their asset allocation studies. Brière et al. (2015) and Kajtazi and Moro (2019) explored the effects of adding Bitcoin to a portfolio with traditional financial assets (worldwide stocks, bonds, and hard currencies) and alternative investments (commodities, hedge funds, and real estate). Additionally, some scholars considered more cryptocurrencies besides Bitcoin such as Ethereum and Ripple in their portfolios with traditional financial assets (see Bouri et al., 2017a; Bouri et al., 2017b; Brauneis and Mestel, 2018; Klein et al., 2018; Trimborn et al., 2020). The above studies confirmed that cryptocurrencies can not only help improve the return/risk ratio of a portfolio, but also diversify the risk of a portfolio. Brauneis and Mestel (2019) used daily data of the 500 most capitalized cryptocurrencies and compared the out-of-sample performance with the mean–variance and 1/N portfolio strategies. The results show that the 1/N portfolio strategy outperforms a single cryptocurrency and more than 75 % of mean–variance portfolio strategies in terms of the Sharpe ratio and certainty equivalent return. Liu (2019) examined the portfolio of 10 cryptocurrencies, and the empirical results suggest that a diversified portfolio of different cryptocurrencies can

* Corresponding author.
E-mail address: Szy_atuo@163.com (Z. Song).

https://doi.org/10.1016/j.eswa.2023.119600
Received 30 June 2022; Received in revised form 4 January 2023; Accepted 21 January 2023
Available online 25 January 2023
0957-4174/

significantly improve investment returns. Also, in agreement with the work of Brauneis and Mestel (2019), no model consistently outperforms the 1/N portfolio in terms of Sharpe ratio, that the out-of-sample performance of traditional mean–variance portfolio strategies is not robust. In addition to discussing the performance of traditional portfolio theory on cryptocurrency assets, many researchers have employed some improved portfolio optimization models based on traditional models. Giudici et al. (2020) combined random matrix theory and network measures with the traditional mean–variance model to construct a new cryptocurrency portfolio model, which outperforms several alternative models with the selected sample data. Aljinović et al. (2021) proposed a multicriteria approach for the cryptocurrency portfolio selection that considers multiple evaluation indicators of cryptocurrencies without limiting to the commonly used return and risk indicators.

The previous cryptocurrency portfolio studies have mainly focused on historical data such as cryptocurrency returns. However, the historical returns of financial assets cannot fully describe the future statistical characteristics of the assets, which induces the out-of-sample performance of the portfolio based on historical returns is often unsatisfactory. With the increasing availability of web-based data and the development of machine learning, deep learning and text mining, investors can forecast the price movements of cryptocurrencies based on multiple data sources. And some scholars also argued that the addition of stock prediction information can improve the performance of portfolios (see Ustun and Kasimbeyli, 2012; Zhou et al., 2021). Therefore, we consider that future price movements in cryptocurrency assets can also be incorporated into constructing the portfolio of cryptocurrencies. Moreover, Liu and Tsyvinski (2021) suggested that cryptocurrency returns can be predicted. However, how to effectively forecast the price movements of cryptocurrencies is an urgent problem to be addressed in this paper.

As far as we know, scholars have utilized many different models to forecast the movement of cryptocurrency prices, which are mainly divided into traditional time series models (see, e.g., Auto Regressive Integrated Moving Average (ARIMA) and Generalized Auto Regressive Conditional Heteroskedasticity (GARCH)), machine learning models (see, e.g., Support Vector Machines (SVM), XGBOOT, and Long Short-Term Memory (LSTM)) (see Adcock and Gradojevic, 2019; Sun et al., 2020; Yun and Shin, 2020; Almansour et al., 2021; Wu, 2021). However, the input feature data of the forecasting model largely determines the model's validity, so selecting the most effective feature data is crucial to improve the accuracy of the forecasting results.

The existing studies have focused on forecasting cryptocurrencies' future prices based on historical trading data (e.g., opening price, closing price, highest price, lowest price, volume, etc.) and media data. Due to the relative youth of the cryptocurrency market, traditional news media outlets cannot always report on cryptocurrency events promptly, which has led to social media becoming the primary source of information for cryptocurrency investors. As one of the world's most popular social media platforms, Twitter provides real-time updates on cryptocurrencies and is a rich source of sentiment data. Moreover, academics increasingly recognize Twitter's ability to predict various events, especially in financial markets (e.g., Arias et al., 2013, Nofer and Hinz, 2015, Oliveira et al., 2017, Mehlawat et al., 2021). In fact, the relationship between cryptocurrencies and investor sentiment on Twitter has been studied extensively in previous research. Kraaijeveld and De Smedt (2020) explored the predictive power of public sentiment on Twitter on the price returns of Bitcoin, Ethereum, Ripple, and other cryptocurrencies using Granger causality tests and sentiment analysis based on cryptocurrency-specific vocabulary. The results show that sentiment messages on Twitter have strong predictive power for Bitcoin, Litecoin, etc. Naseer et al. (2022) extracted tweets related to cryptocurrencies from Twitter, explored the correlation between cryptocurrency price movements and Twitter sentiment, then proposed a method combining deep learning and convolutional neural networks for cryptocurrency forecasting. The results show that Twitter sentiment can improve forecasting accuracy. Aharon et al. (2022) suggested a strong causal relationship between uncertainty expressed in social media and cryptocurrency prices. In conclusion, the studies mentioned above show that investor sentiment on social media can influence the movement of cryptocurrency prices and provide an essential basis for forecasting cryptocurrency prices.

In addition to investor sentiment on social media, investor attention is also an essential factor affecting cryptocurrency price volatility (see Liu and Tsyvinski, 2021). Google, as the most popular search engine on the web today, Google Trends can broadly reflect how much attention people pay to something. Many academics have already used Google Trends to forecast financial assets. Vlastakis and Markellos (2012) demonstrated that Google Trends is an essential driver of future volatility in financial assets (see also Hamid and Heiden, 2015; Andrei and Hasler, 2015). Bijl et al. (2016) used Google Trends to predict stock returns and constructed a trading strategy based on the level of Google search volume. The results show that this strategy is profitable without considering the transaction costs. Qadan and Nama (2018) used daily search query data from Google Trends to determine the attention of retail investors due to oil shocks. The results show that increased searches predicted volatility on subsequent trading days. Liang et al. (2020) used the GARCH-MIDAS model to explore the predictive effect of five predictor variables, including Google Trends, VIX, and GVZ, on bitcoin volatility. The results show that Google Trends has a significant impact on the prediction of bitcoin volatility. In addition, related studies include Ma et al. (2019), Deb (2021), and Dey et al. (2022). Based on the above studies, we believe that Google Trends and investor sentiment on Twitter can be important for forecasting the future movement of cryptocurrency prices.

Based on the above research, this study aims to construct a novel cryptocurrency portfolio optimization model based on the traditional portfolio theory and cryptocurrency price movement forecasting. First, we collect multi-source data of cryptocurrency (i.e. historical trading data, tweets data, and Google Trends). Second, we classify the tweets into sentiment categories with Valence Aware Dictionary and Entiment Reasoner (VADER) and construct the corresponding sentiment indicators. Third, we use SVM to forecast the movement of cryptocurrency prices based on the above multi-source data. Fourth, we propose a novel portfolio model by combining the forecasting results with the global minimum variance model, and the corresponding portfolio strategy is also derived. Finally, we select seven representative cryptocurrencies as the test sample and perform the out-of-sample test of the proposed portfolio strategy and several typical portfolio strategies. The empirical results show that the proposed multi-source data can effectively forecast the movement of cryptocurrency prices; and the proposed strategy has a better out-of-sample Sharpe ratio, Sortino ratio, and Certainty-equivalent (CEQ) return than other strategies. More importantly, the robustness test further confirms the effectiveness of the proposed strategy. Finally, we discuss the case where there are transaction costs, and the results show that the strategy proposed in this paper outperforms other strategies even with transaction costs.

The remainder of this paper is divided into the following sections: Section 2 describes the multi-source data, sentiment analysis method, and the forecasting model used in this paper, and then proposes the cryptocurrency portfolio optimization model. Section 3 selects representative cryptocurrencies as the underlying assets to evaluate the proposed portfolio strategy within the out-of-sample test., then we perform a robustness test to verify our conclusions further, in addition, we discuss the out-of-sample performance with transaction costs. Section 4 provides the conclusion and suggestions of this paper.

2. Methodology

2.1. The description of multi-source data

The first step in forecasting cryptocurrency price movements and constructing a portfolio is to find valuable data related to price changes of a specific cryptocurrency. Based on the above literature, we use historical trading data and social media data to forecast the future movement of cryptocurrency prices. Moreover, we choose tweets on Twitter and Google Trends to represent social media data.

2.1.1. Historical cryptocurrency trading data

We collect historical transaction data of cryptocurrencies, including opening price, closing price, highest price, lowest price, and trading volume. These historical trading data can reflect the historical performance of cryptocurrencies and provide investors with a degree of strategic guidance accordingly.

2.1.2. Social media data

Behavioral economics shows that sentiments can profoundly influence the behavior and decisions of individual investors. With the development of networks, investors often express their views on the underlying assets on online social media platforms, and the social media data can reflect investors' investment tendencies. Several scholars have used social media data to predict cryptocurrency price movements. Furthermore, they have demonstrated the impact of social media data on cryptocurrency price movements (e.g., Kraaijeveld and De Smedt, 2020; Zhang et al., 2021; Naseer et al., 2022).

To the best of our knowledge, as mainstream social media data, Twitter data, and Google Trends have a good performance in forecasting cryptocurrency price movements. On the one hand, the tweet about cryptocurrencies on Twitter can be seen as a bullish or bearish idea about the cryptocurrencies by investors and their willingness to buy or sell the cryptocurrencies; on the other hand, Google Trends on cryptocurrencies represent the search volumes for cryptocurrencies and reflect the attention of investors worldwide. Therefore, we use Twitter data and Google Trends to improve the accuracy of cryptocurrency price movement forecasting. To this end, we crawl tweets on Twitter that contain specific cryptocurrency keywords within a certain period, including tweet time, author, tweet number, and the content of the tweet. In addition, we also collect daily Google Trends for cryptocurrencies to reflect how daily global users pay attention to these cryptocurrencies.

2.2. Tweets sentiment analysis

2.2.1. Data pre-processing

The next problem is how to extract investor sentiment from tweet text data for quantitative analysis. Since the tweet data is unstructured text data, it is hard to be recognized by the programming language without pre-processing and cannot be directly used as model input. Therefore, it is necessary to pre-process the original tweets before conducting sentiment analysis to achieve critical information retention while improving the efficiency of sentiment classification. The detailed steps for pre-processing tweet data are as follows:

- First, filter the tweets data. Twitter users are spread worldwide, and the language used by users is not uniform around the world. A small number of these non-English tweets need to be filtered out. Only the English tweets are subsequently analyzed for the sentiment.
- Second, due to the unstructured nature of tweets, it is necessary to remove useless information from tweets, such as "URL" links, extra spaces, usernames appearing in tweets (e.g., "@account"), "RTs" in retweets, and hashtags with "#" in tweets (e.g., "# Cryptocurrency").

2.2.2. Sentiment analysis using VADER scoring

To categorize tweets, we need to identify and generalize the sentiment words in the tweets and thus determine the sentiment expressed in the tweets towards cryptocurrencies. Generally speaking, sentiment analysis converts text data into a sentiment score representing the sentiment, which is generally classified as positive, negative, or neutral according to the level of the score. In particular, a positive sentiment represents a bullish attitude towards cryptocurrencies, while a negative sentiment represents a bearish attitude towards cryptocurrencies. In this paper, we use the Valence Aware Dictionary and Entiment Reasoner (VADER) algorithm, which performs well in Twitter sentiment analysis for sentiment extraction of tweets (see Chandrasekaran et al., 2020; Ibrahim, 2021).

VADER is a dictionary and rule-based sentiment analysis tool proposed by Hutto and Gilbert (2014), which can handle the analysis of common words, internet abbreviations, slang, face characters and emojis, etc., in social media. More importantly, VADER is much faster than standard machine learning algorithms because it is used directly without training. For each text, it generates a vector of sentiment scores with negative, neutral, positive, and compound polarity. Compound score is considered as a composite measure of all other emotions and are the most useful metric, which is normalized between -1 (negative) and 1 (positive). Hutto and Gilbert (2014) also compared VADER with other sentiment analysis models. Their results show that VADER performs better than most other sentiment analysis models for Twitter text analysis. Therefore, we use the compound score to determine investor sentiment towards cryptocurrencies as expressed in tweets. And we also categorize tweets into three categories $c \in \{pos, neu, neg\}$, where *pos* means positive sentiment (compound score $>= 0.05$), *neg* means negative sentiment (compound score $<= 0.05$), and *neu* means neutral sentiment (compound score > -0.05 and compound score < 0.05).

2.2.3. Constructing sentiment indicators

In this part, we construct investor sentiment indicators based on tweets' sentiment classification results to reflect better the information on investor sentiment embedded in tweets. Following Antweiler and Frank (2004), Tsukioka et al. (2018), Li et al. (2020), and Kraaijeveld and De Smedt (2020), we adopt three sentiment indicators,B_t, B_t^* and B_t^{**}, B_t is defined as Eq.(1):

$$B_t = \frac{M_t^{pos} - M_t^{neg}}{M_t^{pos} + M_t^{neg}}, \tag{1}$$

where $M_t^c = \sum_{i \in D(t)} w_i x_i^c$ denotes the weighted sum of messages of type $c \in \{pos, neu, neg\}$ in the time interval $D(t)$, where x_i^c is an indicator variable equal to 1 when the tweet i is the type c and zero otherwise, and w_i is the weight of the tweets. Antweiler and Frank (2004) showed that changing the weighting scheme did not affect their results and used the equal weighting method. Therefore, following Antweiler and Frank's work, we let $w_i = 1$, i.e., M_t^c denotes the total number of tweets with sentiment type c per day. Antweiler and Frank define another indicator, as shown in Eq.(2):

$$B_t^* = \ln\left[\frac{1 + M_t^{pos}}{1 + M_t^{neg}}\right], \tag{2}$$

and the relationship between these two indicators is $B_t^* \approx B_t \ln\left(1 + \left(M_t^{pos} + M_t^{neg}\right)\right)$. The B_t^* considers the number of tweets that express a particular sentiment, whereas Eq.(1) only considers the difference in the number of negative and positive sentiment, so B_t^* is superior to the other alternatives.

Since the two sentiment indicators B_t and B_t^* proposed by Antweiler and Frank (2004) above contain only positive and negative sentiment information and do not consider neutral sentiment information. Even though sentiment-neutral tweets do not reflect positive or negative investor sentiment, they help get investors' attention, so neutral information is also valuable. Following Li et al. (2020), we combine sentiment-neutral tweets to construct the third sentiment indicator B_t^{**}, as shown in Eq.(3):

$$B_t^{**} = B_t \ln(1 + M_t), \tag{3}$$

where M_t denotes the total amount of tweets at the time interval $D(t)$ expressing investor attention to information about cryptocurrencies. M_t changes with the investor attention on cryptocurrencies but is not affected by the sentiment analysis algorithms. In this paper, the daily investor sentiment for day t is extracted according to the tweets from 00:00 to 24:00 of the current day t. In addition to the sentiment indicators B_t, B_t^* and B_t^{**}, we also use daily Google Trends to represent investor sentiment and attention to the cryptocurrency market.

2.3. Cryptocurrency forecasting models

Support Vector Machine (SVM) is a powerful binary classification machine learning algorithm first proposed by Cortes and Vapnik (1995) based on the principle of structural risk minimization. The aim is to classify-two classes by finding an optimal hyperplane that maximizes the spacing between points in the multidimensional space. SVM is outstanding in handling classification, identification, and regression problems. Therefore, based on the above multi-source data, the SVM algorithm can be used to forecast the future price movements of cryptocurrencies (Mahdi et al., 2021). Fig. 1 shows the framework for forecasting cryptocurrencies.

Referring to the previous research, such as Huang et al. (2005), Burges (1998), Mahdi et al. (2021), and Wang et al. (2019), we briefly describe the idea of the SVM algorithm as follows:

(1) Given a training set $\{(x_i,\ y_i),\ i=1,\ 2,\ \cdots,\ n\}$, where $x_i \in \mathbb{R}^d$, $y_i \in \{-1,1\}$, n is the number of training samples (in this paper, n denote the number of trading days).

(2) To classify-two different classes of data points, the SVM finds an optimal hyperplane by solving the following quadratic optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \qquad (4)$$
$$s.t.\ y_i\left[w^T\varphi(x_i)+b\right]+\xi_i-1\geqslant 0,\ and\ \xi_i\geqslant 0, i=1,2,\cdots,n.$$

Where $w \in \mathbb{R}^d$, C is the penalty parameter and $C>0$, ξ_i is the non-negative slack variable, and $\varphi(x_i)$ is the feature mapping function, which maps the input space to a higher-dimensional feature space where the data points become linearly separable. In addition, the dual form of Model (4) can be expressed as follows:

$$\max_{\alpha i} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \varphi(x_i)^T\varphi(x_j), \qquad (5)$$
$$s.t.\ 0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,n, and \sum_{i=1}^{n}\alpha_i y_i = 0.$$

Here, $\alpha_i\geqslant 0,\ i=1,\ 2,\ \cdots,\ n$, are the Lagrange multipliers. In the dual Lagrangian Model (5), the inner products can be replaced by a kernel function K, i.e., $K(x_i,x_j)=\varphi(x_i)^T\cdot\varphi(x_j)$, there are four general types of kernel functions commonly used, which are linear function, polynomial function, radial basis function, and Sigmoid function.

Note that $\alpha_i^* = (\alpha_1^*,\alpha_2^*,\cdots,\alpha_n^*)$ denotes the solution to the dual problem and b^* follows the complementarity KKT conditions, combined with Model (5), the decision function can be written in the following form:

$$f(x) = sgn\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x) + b^*\right). \qquad (6)$$

Let x denotes the input vector, and x_i denotes the i-th learning sample point, the connection weight of the input layer to the hidden layer is 1, $K(x,\ x_i)$ denotes the connection right of the i-th neuron of the hidden layer to the neuron of the output layer. In addition, $y=f(x)$ represents the output of SVM. Moreover, the flow of the above SVM classification algorithm is shown in Fig. 2:

2.4. The proposed portfolio optimization model

The existing portfolio models typically construct optimization objectives based on the return and risk of assets and then use historical returns as the sample to estimate the parameters of portfolios. However, the historical return characteristics of financial assets often do not reflect the future statistical characteristics of the underlying assets, and that makes the out-of-sample performance often unsatisfactory (e.g., Michaud, 1989; DeMiguel et al., 2009; Tu and Zhou, 2011). Therefore, to solve and quantify portfolio optimization in the cryptocurrency field, we combine forecasts of future movement of cryptocurrency prices with traditional portfolio theory to provide cryptocurrency investors with a novel portfolio strategy. Since the existing literature has shown that the

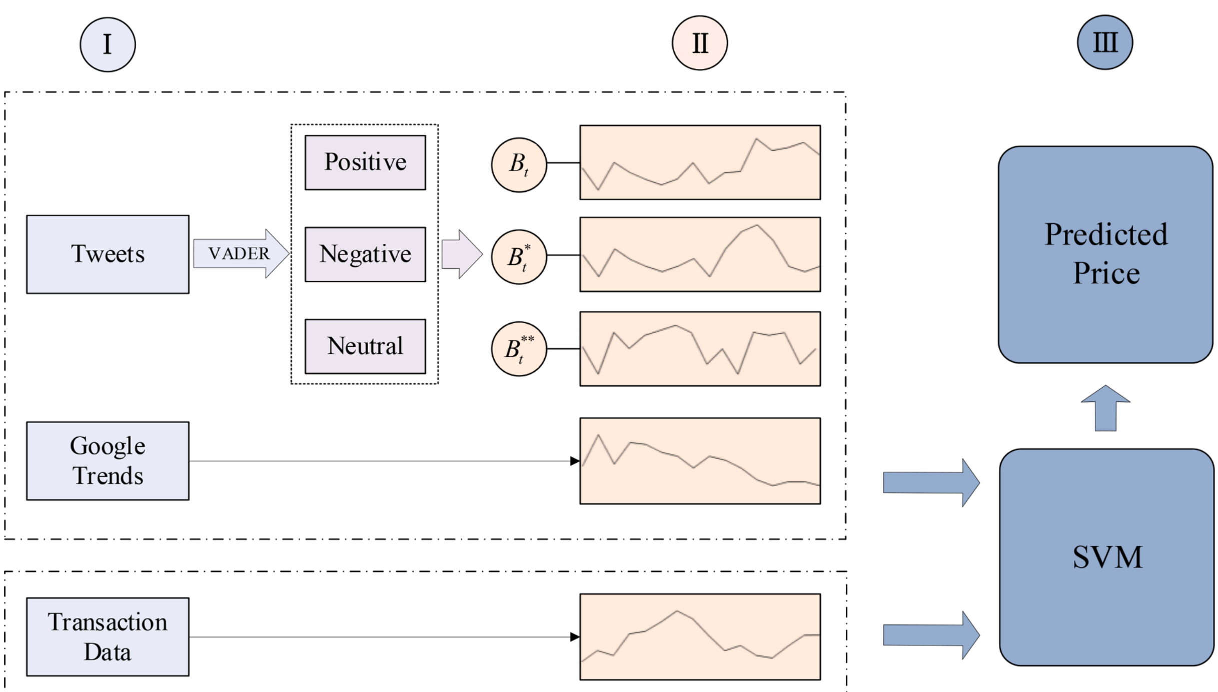

Fig. 1. The framework for forecasting the movement of cryptocurrency prices.

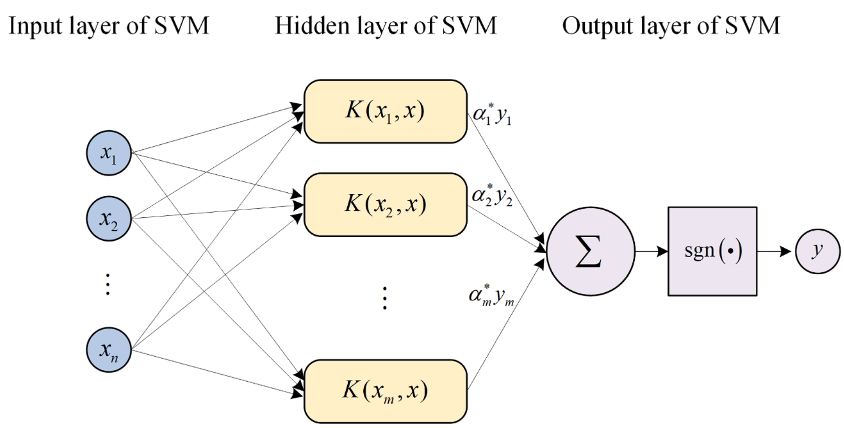

Fig. 2. Support vector machine classification algorithm.

minimum variance portfolio usually has better out-of-sample performance than other traditional portfolios (e.g., Chopra and Ziemba, 1993; Jagannathan and Ma, 2003; Ledoit and Wolf, 2003). Thus, we combine the minimum variance portfolio model with cryptocurrency forecasting information and propose a novel cryptocurrency portfolio optimization model, and the model is shown below:

$$\begin{aligned} &\min_x x'\Omega_n x - \phi \times x' f_n, \\ &s.t.\, n^* \sum_{i=1}^{n} x_i = 1 \times n^*, x_i \geqslant 0, i = 1, 2, \cdots, n. \end{aligned} \tag{7}$$

Here, x denotes the investment weight matrix for each asset, Ω_n denotes the covariance matrix of the asset, and $f_n = (f_n^1, f_n^2, \cdots, f_n^n)'$ is the forecasting vector for the rise and fall of cryptocurrencies prices. f_n^i is a 0–1 variable, where $f_n^i = 1$ denotes the price of the cryptocurrency i will not fall in the future, and $f_n^i = 0$ denotes the price of the cryptocurrency i will fall in the future. The parameter ϕ is a predetermined constant parameter that can be used as a trade-off parameter between the cryptocurrency price forecasting results and the minimum variance portfolio. n^* represents the number of predicted values of 1 in f_n, and $n^* = \sum_{i=1}^{n} I' f_n, I = (1, 1, \cdots, 1)'_{1\times n}, n^* \geqslant 0$, when the prices of all cryptocurrencies are predicted to fall, i.e. $n^* = 0$, it is obvious the optimal solution x^* of the model is always 0, which means when we predict the prices of all underlying assets will fall in the future, we do not need to make any investments. To make the proposed model better understood, we will illustrate it with an example: If we want to invest in 5 cryptocurrency assets today, we can use SVM to predict the price movement of these 5 assets in tomorrow, assume that the obtained prediction matrix is $f_n = (1, 0, 0, 0, 0)'$, means that we predict that the price of the first asset will rise tomorrow, while the price of the other assets will fall. At this point, since our optimizer is a minimization function, the model will want the asset with a prediction of 1 to have as much of the investment share as possible. Therefore, the effectiveness of the portfolio can be improved by forecasting information. And when $n^* > 0$, Model (7) is expressed as follows:

$$\begin{aligned} &\min_x x'\Omega_n x - \phi \times x' f_n, \\ &s.t. \sum_{i=1}^{n} x_i = 1, x_i \geqslant 0, i = 1, 2, \cdots, n. \end{aligned} \tag{8}$$

the corresponding portfolio strategy is obtained, denoted as $x^{Sentiment}$.

In addition, following DeMiguel et al. (2009), Brauneis and Mestel (2019), and Liu (2019), compared to other traditional portfolio strategies, the 1/N portfolio has better performance at some times. Therefore, we build another portfolio strategy based on the 1/N portfolio strategy and cryptocurrency future price forecasts, and compare it with Model (7). The specific portfolio strategy is shown below:

$$x^{Sentiment-ew} = \begin{cases} \frac{1}{n^*} f_n, n^* > 0, \\ 0, n^* = 0. \end{cases} \tag{9}$$

For convenience, Strategy (9) is denoted as $x^{Sentiment-ew}$.

The proposed portfolio strategy combines cryptocurrency forecasting results with the minimum variance theory. Considering the historical returns and the future forecasting information of cryptocurrencies enables a trade-off between past and future performance, largely alleviating the previous strategy's over-reliance on the historical performance of the underlying assets.

In order to demonstrate the effectiveness of the proposed portfolio strategies, we first introduce some classical portfolio strategies for comparison. First, disregarding cryptocurrency forecasts and focusing only on historical asset returns, we construct the minimum variance portfolio model as follows:

$$\begin{aligned} &\min_x x'\Omega_n x, \\ &s.t. \sum_{i=1}^{n} x_i = 1, x_i \geqslant 0, i = 1, 2, \cdots, n \end{aligned} \tag{10}$$

For convenience, the minimum variance portfolio strategy is denoted as x^{mv}.

Second, we consider the tangency Portfolio (Maximum Sharpe ratio portfolio) to compare the proposed strategy. The tangency portfolio optimization model is shown below:

$$\begin{aligned} &\max_x \frac{R_n - r_f}{\sqrt{x'\Omega_n x}}, \\ &s.t. \sum_{i=1}^{n} x_i = 1, x_i \geqslant 0, i = 1, 2, \cdots, n. \end{aligned} \tag{11}$$

Here, R_n is the mean vector of historical returns and $R_n = (r_1, r_2, \cdots, r_n)'$. For convenience, the tangency portfolio strategy is denoted as $x^{\max}$.

Then, we introduce the 1/N portfolio strategy for comparison with the proposed strategy, and the 1/N model is specified as follows:

$$x^{ew} = \frac{1}{n} I. \tag{12}$$

For convenience, the 1/N portfolio strategy is denoted as x^{ew}, where $I = (1, 1, \cdots, 1)'_{1\times n}$.

In addition to the above traditional portfolio obtained based on historical information, we also introduce some other portfolio strategies as benchmarks. We first consider the Black-Litterman approach which incorporate investors' views when constructing portfolio strategies. To construct investors' views, we follow Sahamkhadam et al. (2022) use the Vector Error Correction Model (VECM) to create expectations of price changes. For convenience, the BL portfolio strategy is denoted as x^{bl}.

We also chose a approach that uses machine learning methods to predict assets for portfolio investment as a comparison. Li et al (2019) used historical data and technical indicators of assets to predict the return of assets using LSTM, then ranked the assets by the predicted returns, and preselecting them according to the ranking. And the top assets are selected for investment according to the minimum variance portfolio strategy. Referring to Li et al., the best results were obtained by selecting 3 assets for investment, so we use historical data and technical indicators to predict the future return of cryptocurrencies and rank the assets according to the obtained returns, and select the top 3 assets as assets pool, then using the minimum variance portfolio model to obtained investment weights. For convenience, the portfolio strategy proposed by Li et al. is denoted as $x^{Li\ et\ al.}$.

In addition, this paper also introduces the CRyptocurrency IndeX (CRIX) proposed by Trimborn and Härdle (2018) to compare the above portfolio strategies. CRIX is a market capitalization-weighted index commonly used to reflect the overall cryptocurrency market. For convenience, we give all the portfolio strategies considered and the corresponding abbreviations, as seen in Table 1.

3. Empirical analysis

Based on the market capitalization, we select seven representative cryptocurrencies as the research objects for the future movement of cryptocurrency prices forecasting and for measuring the out-of-sample performance of the proposed approach, including Bitcoin, Ethereum, Ripple, Cardano, Dogecoin, Polkadot, and Litecoin. Table 2 depicts the market characteristics of these cryptocurrencies as of February 14, 2022. First, we collect daily historical trading data (i.e., opening price, closing price, high price, low price, and trading volume) for cryptocurrencies from cn.investing.com from January 1, 2021 to June 30, 2021. Second, we use TweetScraper in Python to crawl tweets from Twitter related to each cryptocurrency from January 1, 2021, to June 30, 2021, and collect 28,799,774 tweets. The number of collected tweet data is shown in Table 3. Then, we use VADER to analyze the sentiment of each tweet, categorizing it into positive, negative, and neutral sentiment. Fig. 3 shows the percentage of tweets for different sentiment types for cryptocurrencies after sentiment analysis. After sentiment analysis, we can calculate the daily sentiment indicators for the tweets data. In addition to the tweets data, we also collect the daily Google Trends of cryptocurrencies from https://trends.google.com.

3.1. Cryptocurrency price forecasting

This section constructs vector matrices with multi-source data (i.e., historical trading data of cryptocurrencies, sentiment indicators, and Google Trends) as input data for the SVM algorithm. The future movement of cryptocurrencies prices are output data. The price movement is the difference between the closing prices of two trading days, which can be defined as $Close(t+\tau) - Close(t)$. Furthermore, if $Close(t+\tau) - Close(t) > 0$, that means the cryptocurrency price is rise after τ days, i.e., the value of cryptocurrency price movement $\hat{y} = 1$, otherwise $\hat{y} = 0$. Previous study has typically considered the performance of investment strategies under only one holding period and ignored the performance of the portfolio strategies under different holding periods (see Zhou et al., 2021). To compare the out-of-sample performance between the proposed portfolio strategy and the traditional portfolio strategies under different holding periods, we consider holding cryptocurrency assets for one day, three days, and five days, respectively. Therefore, it is necessary to forecast the price movements after 1, 3, and 5 days, i.e.,$\tau = 1$, $\tau = 3$ and $\tau = 5$, respectively.

To measure the results, we need some evaluation metrics. The confusion matrix is the most commonly used measures to determine the quality of the SVM binary classification forecasting method. The confusion matrix provides a visual performance evaluation of the classification algorithm in the form of a matrix, which can be used to evaluate the classification forecasting results. The details are shown in Table 4, where TP represents Ture Positives, FP represents False Positives, TN represents Ture Negatives, and FN represents False Negatives. (see Niu et al., 2008; Fayyaz et al., 2020; Ibrahim, 2021). We use Accuracy as the primary criterion for evaluating the performance of SVM models. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}. \tag{13}$$

Since accuracy is not always reliable, this paper introduces Precision, Recall, and F1-score as additional metrics to evaluate the performance of these models. The formulas for calculating these metrics are shown below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{14}$$

$$Recall = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{15}$$

$$\text{F1}-\text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}\ +\ \text{Recall}}. \tag{16}$$

First, we arrange the collected data in chronological order while normalizing the input data set, and then use the SVM classification algorithm to train and forecast the input data.

Table 1
List of various portfolio models considered.

#	Model	Abbreviation
1.	Minimum variance with forecasting information	$x^{Sentiment}$
2.	1/N with forecasting information	$x^{Sentiment-ew}$
3.	Minimum variance strategy	x^{mv}
4.	Maximum Sharpe ratio strategy (Tangency strategy)	x^{max}
5.	1/N strategy (Equal-weight strategy)	x^{ew}
6	Li et al.' s (2019) Preselection approach	$x^{Li\ et\ al.}$
7	Black-Litterman approach	x^{bl}
8.	CRyptocurrency IndeX	*CRIX*

Table 2
Characteristics of the cryptocurrencies (February 14, 2022). Source: https://www.binance.com/en/markets/coinInfo.

Cryptocurrency	Price (in US $)	24 h Volume(in US $)	Market Cap (in US $)
Bitcoin(BTC)	42,559.85	18,394.78 million	807.295.13 million
Ethereum(ETH)	29,932.48	10,424.20 million	350,644.85 million
Ripple(XRP)	0.08045	2,658.97 million	38,390.35 million
Cardano(ADA)	1.05000	1,273.07 million	35,272.27 million
Dogecoin (DOGE)	0.14670	1,733.49 million	19,449.53 million
Polkadot(DOT)	18.7300	824.51 million	18,517.48 million
Litecoin(LTC)	126.5000	753.15 million	8,815.98 million

Table 3
The number of collected Tweets.

Cryptocurrency	Total number of collected Tweets
Bitcoin(BTC)	12550256
Ethereum(ETH)	5292492
Ripple(XRP)	3098863
Cardano(ADA)	867121
Dogecoin(DOGE)	5718898
Polkadot(DOT)	379502
Litecoin(LTC)	892642
Total	28799774

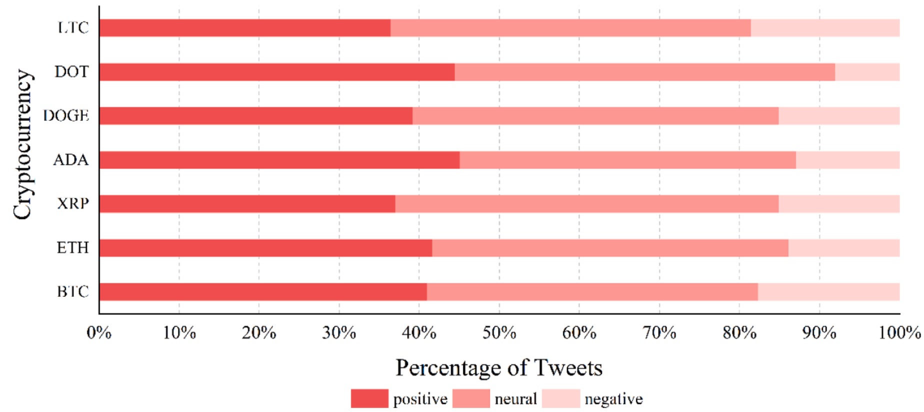

Fig. 3. Percentage of tweets for different sentiment types.

Table 4
The details of Confusion Matrix.

Actual value \ Predicted value	1	0
1	TP	FN
0	FP	TN

Then, we use 6 months of data from January 1, 2021, to June 30, 2021 as the dataset for the SVM. And we predict the future price movements of each cryptocurrency on the next day, after three days, and after five days, respectively. The first 80 % of data is the training set, and the last 20 % is the test set. Moreover, according to the test set results, using the Grid Search method for model parameters tuning. The next-day price forecasts, three-day price forecasts, and five-day price forecasts for cryptocurrencies at $\tau = 1$, $\tau = 3$ and $\tau = 5$ are obtained by adjusting the parameters. The accuracy, recall, precision, and F1-score of the cryptocurrencies test sets are shown in Table 5. Furthermore, to test whether the proposed multi-source data can effectively improve forecasting performance, we also calculate the case of considering only historical trading data, as shown in parentheses.

Table 5
The results of cryptocurrency forecasting.

Cryptocurrency	Accuracy	Precision	Recall	F1-score
The case where $\tau = 1$				
BTC	0.61 (0.58)	0.65 (0.58)	0.61 (0.58)	0.60 (0.58)
ETH	0.61 (0.53)	0.63 (0.28)	0.61 (0.53)	0.60 (0.36)
XRP	0.72 (0.61)	0.81 (0.62)	0.72 (0.61)	0.66 (0.62)
ADA	0.58 (0.53)	0.58 (0.52)	0.58 (0.53)	0.58 (0.51)
DOGE	0.72 (0.62)	0.81 (0.61)	0.72 (0.61)	0.66 (0.62)
DOT	0.67 (0.58)	0.80 (0.67)	0.67 (0.58)	0.62 (0.52)
LTC	0.61 (0.50)	0.64 (0.45)	0.61 (0.50)	0.59 (0.33)
Average	0.65 (0.56)	0.70 (0.53)	0.65 (0.56)	0.62 (0.51)
The case where $\tau = 3$				
BTC	0.59 (0.53)	0.63 (0.76)	0.59 (0.53)	0.57 (0.42)
ETH	0.62 (0.50)	0.60 (0.79)	0.62 (0.50)	0.60 (0.45)
XRP	0.65 (0.56)	0.64 (0.64)	0.65 (0.56)	0.64 (0.56)
ADA	0.59 (0.39)	0.73 (0.35)	0.59 (0.32)	0.57 (0.28)
DOGE	0.71 (0.47)	0.76 (0.34)	0.71 (0.47)	0.71 (0.40)
DOT	0.68 (0.44)	0.67 (0.77)	0.68 (0.41)	0.64 (0.33)
LTC	0.62 (0.56)	0.65 (0.55)	0.62 (0.56)	0.62 (0.55)
Average	0.63 (0.49)	0.67 (0.60)	0.64 (0.48)	0.62 (0.43)
The case where $\tau = 5$				
BTC	0.59 (0.53)	0.59 (0.52)	0.59 (0.53)	0.59 (0.42)
ETH	0.63 (0.38)	0.61 (0.78)	0.62 (0.38)	0.61 (0.24)
XRP	0.75 (0.63)	0.75 (0.39)	0.75 (0.62)	0.75 (0.48)
ADA	0.59 (0.44)	0.61 (0.56)	0.59 (0.44)	0.60 (0.34)
DOGE	0.72 (0.63)	0.73 (0.46)	0.72 (0.62)	0.72 (0.53)
DOT	0.59 (0.41)	0.56 (0.14)	0.59 (0.38)	0.56 (0.20)
LTC	0.75 (0.47)	0.75 (0.77)	0.75 (0.47)	0.75 (0.36)
Average	0.66 (0.50)	0.66 (0.52)	0.66 (0.49)	0.65 (0.37)

From the above results, it is clear that the proposed multi-source data can help forecast the future price movements of cryptocurrency. It is also demonstrated that the cryptocurrency forecasting method using multi-source data and SVM can better forecast the future price movements of cryptocurrencies, which validates the findings of Kraaijeveld and De Smedt (2020). This further demonstrates that it is plausible to use cryptocurrency price prediction results to construct portfolio strategies.

3.2. Out-of-sample tests of portfolio strategies

Based on the forecasting results in Section 3.1 and the historical return data, we use the "rolling sample" approach mentioned by Demiguel et al. (2009) to perform the out-of-sample test of the cryptocurrencies portfolio strategies in Section 2.4. The test interval is from January 1, 2021 to June 30, 2021, and we assume the length of the window in the "rolling sample" is M, and the parameters $\phi = 0.001$ in Model (7).

Let May 26, 2021, as the starting point (i.e., $t = 1$). The historical return data (noted as r_c) for M days before that point is used as the sample data for the benchmark and parameter estimation, and we calculate the covariance matrix Ω_n between cryptocurrencies and the price forecasts f_n. Further, the portfolio strategies shown in Section 2.4 is derived. The corresponding portfolio return per holding period r_p is calculated based on the realized return for the next trading period.

Similarly, when $t = 2$, we can update the covariance matrix Ω_n between cryptocurrencies by using the return in the previous M days. In addition, the price forecasts f_n and cryptocurrency investment strategies can also be updated. Then, we can also calculate the portfolio return r_p for the next trading period. This part discusses the out-of-sample performance of different investment strategies with a 1-day, 3-day, and 5-day holding period, respectively.

By using the rolling window method, at different holding periods, we can obtain a series of 36, 34, and 32 out-of-sample portfolio returns for each strategy, respectively. Therefore, the out-of-sample performance of each portfolio strategy is evaluated by the following four evaluation metrics.

First, we introduce the *Sharpe ratio* as one of the metrics for evaluating the performance of different portfolio strategies, which is defined as:

$$\theta^{Sharpe} = \frac{E(r_p) - r_f}{\sqrt{Var(r_p)}}. \tag{17}$$

Here, $E(r_p)$ represents the mean of out-of-sample returns for strategy p, and $Var(r_p)$ represents the out-of-sample variance, r_f is the risk-free rate of return and we assume $r_f = 0$.

Second, we introduce the *Sortino ratio* as one of the metrics for evaluating the performance of portfolio strategies. The *Sortino ratio* is defined as:

$$\theta^{Sortino} = \frac{E(r_p) - r_f}{\sqrt{E\left[\left(r_p - E(r_p)\right)_-^2\right]}}. \tag{18}$$

Here, $(R)_- = \min\{0, R\}$.

Third, we calculate the *certainty-equivalent (CEQ) return*, which is defined as the risk-free rate of return that an investor is willing to accept rather than adopting a particular risky portfolio strategy. The *CEQ return* is defined as:

$$\theta^{CEQ} = E(r_p) - \frac{\gamma}{2} Var(r_p). \tag{19}$$

Here, γ is the risk aversion, and we let $\gamma = 1$. Meanwhile, we calculate the *return-loss* between each strategy and the proposed strategy. The *return-loss* is defined as the additional return needed for strategy p to perform as well as the proposed strategy $x^{Sentiment}$ in terms of the Sharpe ratio. And the $E(r_S)$ and $Var(r_S)$ is the mean and variance of out-of-sample returns for the proposed strategy $x^{Sentiment}$, then $E(r_p)$ and $Var(r_p)$ is the mean and variance of out-of-sample returns for strategy p. Therefore, the return-loss from strategy p relative to strategy $x^{Sentiment}$ is defined as:

$$return - loss = \frac{E(r_s)}{\sqrt{Var(r_s)}} \times \sqrt{Var(r_p)} - E(r_p). \tag{20}$$

Based on the above evaluation metrics, following Demiguel et al. (2009), we assume that the window lengths M are 30, 60, 90, and 120, respectively. Then, we calculate the out-of-sample Sharpe ratio, Sortino ratio, CEQ Return, and return-loss for each strategy with a holding period of 1, 3, and 5 days, respectively, The results are shown in Tables 6–14:

When the holding period is one day, and the window length M is respectively at 30, 60, 90, and 120, the results of Sharpe and Sortino ratios are shown in Tables 6–7. The proposed portfolio strategy $x^{Sentiment}$ has a higher out-of-sample Sharpe/Sortino ratio than the strategies $x^{\max}$,x^{ew},$x^{Li\ et\ al.}$,x^{bl}, and *CRIX*. In addition, the proposed strategy $x^{Sentiment}$ outperforms strategy $x^{Sentiment-ew}$ in most cases, further demonstrating that the minimum variance portfolio strategy outperforms the 1/N portfolio strategy in most cases. Although strategy $x^{Sentiment-ew}$ underperforms strategy $x^{Sentiment}$, the out-of-sample performance consistently outperforms the 1/N strategy x^{ew}. The results in Table 6 and Table 7 suggest that the proposed portfolio optimization approach can significantly improve the out-of-sample performance of the portfolio strategies. The comparison of the CEQ return in Table 8 further supports the conclusions drawn from the Sharpe ratio and Sortino ratio analysis. The proposed strategy $x^{Sentiment}$ always has a higher CEQ return than the strategy $x^{Sentiment-ew}$,x^{mv} ,$x^{\max}$,x^{ew},$x^{Li\ et\ al.}$,x^{bl}, and CRIX when the window length $M = 30$,$M = 60$,$M = 90$, and $M = 120$. Also, the results of Panel B further demonstrate that strategy $x^{Sentiment}$ achieves a higher Sharpe ratio than other strategies. Overall, in most cases, the proposed strategy $x^{Sentiment}$ has a higher out-of-sample performance than other portfolio strategies.

From the results in Table 8, Table 9, and Table 10, the proposed portfolio strategy has a higher Sharpe ratio and Sortino ratio compared to most other strategies when the holding period is three days, regardless of the window length $M = 30$,$M = 60$, $M = 90$ or $M = 120$. This finding is consistent with the results in Table 6 and Table 7. It is found that strategy $x^{Sentiment}$ outperforms strategy x^{mv} for different window lengths. In contrast, strategy $x^{Sentiment-ew}$ consistently outperforms strategy x^{ew} in terms of out-of-sample Sharpe ratio and Sortino ratio. This further illustrates the effectiveness and usefulness of incorporating cryptocurrency forecasting information into portfolio construction. It is interesting to note that strategy $x^{Li\ et\ al.}$ performs better than the other strategies at t = 3. The CEQ return in Table 11 further support the conclusions obtained from the Sharpe ratio and Sortino ratio analysis. From Panel B, it can be seen that the proposed strategy $x^{Sentiment}$ consistently outperforms most of the other portfolio strategies in terms of the Sharpe ratio. Overall, the above results show that the proposed portfolio can significantly improve the out-of-sample performance of cryptocurrency portfolios when the holding period is 3 days.

Table 6
The out-of-sample Sharpe ratio of different portfolio strategies ($\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.0499	0.0354	0.0351	0.0284
$x^{Sentiment-ew}$	0.0280	0.0280	0.0280	0.0280
x^{mv}	−0.0326	0.0266	0.0266	−0.0214
$x^{\max}$	−0.1069	0.0753	0.0596	−0.0283
x^{ew}	−0.0582	0.0582	0.0582	−0.0582
$x^{Li\ et\ al.}$	−0.1578	−0.1539	−0.1430	−0.1458
x^{bl}	−0.1064	−0.1411	−0.0215	−0.1895
CRIX	−0.1988	0.1988	0.1988	−0.1988

Table 7
The out-of-sample Sortino ratio of different portfolio strategies ($\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.0709	0.0507	0.0503	0.0405
$x^{Sentiment-ew}$	0.0370	0.0370	0.0370	0.0370
x^{mv}	−0.0458	−0.0384	−0.0384	−0.0309
$x^{\max}$	−0.1499	−0.1064	−0.0811	−0.0386
x^{ew}	−0.1357	−0.1357	−0.1357	−0.1357
$x^{Li\ et\ al.}$	−0.2130	−0.2088	−0.1944	−0.1931
x^{bl}	−0.1501	−0.1778	−0.0342	−0.2287
CRIX	−0.2764	−0.2764	−0.2764	−0.2764

Table 8
The out-of-sample CEQ return and return-loss of different portfolio strategies ($\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	0.0012	0.0005	0.0005	0.0002
$x^{Sentiment-ew}$	0.0001	0.0001	0.0001	0.0001
x^{mv}	−0.0030	−0.0026	−0.0026	−0.0023
$x^{\max}$	−0.0078	−0.0063	−0.0083	−0.0059
x^{ew}	−0.0056	−0.0056	−0.0056	−0.0056
$x^{Li\ et\ al.}$	−0.0112	−0.0102	−0.0092	−0.0097
x^{bl}	−0.0048	−0.0214	−0.0054	−0.0249
CRIX	−0.0155	−0.0155	−0.0155	−0.0155
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	0.0011	0.0004	0.0004	0.0000
x^{mv}	0.0043	0.0031	0.0031	0.0025
$x^{\max}$	0.0090	0.0066	0.0078	0.0047
x^{ew}	0.0068	0.0059	0.0059	0.0054
$x^{Li\ et\ al.}$	0.0124	0.0106	0.0096	0.0098
x^{bl}	0.0060	0.0193	0.0048	0.0225
CRIX	0.0166	0.0156	0.0156	0.0152

Table 9
The out-of-sample Sharpe ratio of different portfolio strategies ($\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	−0.0328	−0.0477	−0.0401	−0.0517
$x^{Sentiment-ew}$	−0.0506	−0.0506	−0.0506	−0.0506
x^{mv}	−0.0654	−0.0746	−0.0688	−0.0902
$x^{\max}$	−0.1625	−0.1618	−0.1539	−0.1064
x^{ew}	−0.1616	−0.1616	−0.1616	−0.1616
$x^{Li\ et\ al.}$	−0.1032	−0.0026	−0.0117	−0.0373
x^{bl}	−0.1348	−0.0811	−0.1262	−0.1047
CRIX	−0.3830	−0.3830	−0.3830	−0.3830

Table 10
The out-of-sample Sortino ratio of different portfolio strategies ($\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	−0.0491	−0.0714	−0.0597	−0.0766
$x^{Sentiment-ew}$	−0.0737	−0.0737	−0.0737	−0.0737
x^{mv}	−0.0995	−0.1132	−0.1041	−0.1352
$x^{\max}$	−0.2665	−0.2606	−0.2308	−0.1625
x^{ew}	−0.3413	−0.3413	−0.3413	−0.3413
$x^{Li\ et\ al.}$	−0.1540	−0.0038	−0.0169	−0.0534
x^{bl}	−0.2269	−0.1322	−0.1917	−0.1541
CRIX	−0.5123	−0.5123	−0.5123	−0.5123

Table 11
The out-of-sample CEQ return and return-loss of different portfolio strategies ($\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	−0.0054	−0.0065	−0.0061	−0.0069
$x^{Sentiment-ew}$	−0.0102	−0.0102	−0.0102	−0.0102
x^{mv}	−0.0081	−0.0086	−0.0083	−0.0097
$x^{\max}$	−0.0191	−0.0220	−0.0283	−0.0217
x^{ew}	−0.0221	−0.0221	−0.0221	−0.0221
$x^{Li\ et\ al.}$	−0.0164	−0.0057	−0.0068	−0.0093
x^{bl}	−0.0123	−0.0121	−0.0192	−0.0178
CRIX	−0.0434	−0.0434	−0.0434	−0.0434
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	0.0018	0.0003	0.0011	−0.0001
x^{mv}	0.0025	0.0021	0.0022	0.0029
$x^{\max}$	0.0119	0.0118	0.0147	0.0070
x^{ew}	0.0133	0.0118	0.0126	0.0114
$x^{Li\ et\ al.}$	0.0074	−0.0047	−0.0030	−0.0015
x^{bl}	0.0073	0.0032	0.0092	0.0059
CRIX	0.0351	0.0336	0.0343	0.0332

In Table 12, Table 13, and Table 14, it can be seen that in the case of $\tau = 5$, the results are different from those in $\tau = 1$ and $\tau = 3$. When $\tau = 5$, the strategy that obtains the highest Sharpe ratio and Sortino ratio is strategy $x^{Sentiment-ew}$. However, the proposed strategy $x^{Sentiment}$ still consistently outperforms the other portfolio strategies and the CRIX in terms of Sharpe ratio and Sortino ratio, suggesting that the forecasts of future movements of cryptocurrencies price are equally effective in improving the out-of-sample performance of the portfolios when the holding period is five days. The results of the CEQ return confirm the conclusion from the Sharpe/Sortino ratio analysis that the portfolio strategies incorporating price forecasting information have higher CEQ rates than other portfolio strategies and CRIX. Moreover, in most cases, $x^{Sentiment}$ has the highest CEQ rates. Also, the results of Panel B show that when the holding period is five days, the proposed strategy $x^{Sentiment}$ can also achieve a higher Sharpe ratio than other traditional strategies.

Furthermore, probably due to the high volatility characteristics of cryptocurrencies themselves, from the results of Tables 6–14, we find that the out-of-sample performance of each strategy tends to decrease as the length of the window M increases, regardless of whether the holding period is 1, 3, or 5 days. This suggests that the historical performance of cryptocurrency assets should not be given too much consideration when constructing a portfolio. Overall, the proposed portfolio strategy can significantly improve the out-of-sample performance when the holding period is 1, 3, and 5 days.

Table 12
The out-of-sample Sharpe ratio of different portfolio strategies ($\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	−0.0601	−0.0698	−0.0481	−0.1027
$x^{Sentiment-ew}$	−0.0434	−0.0434	−0.0434	−0.0434
x^{mv}	−0.0923	−0.0967	−0.0796	−0.1358
$x^{\max}$	−0.1793	−0.2020	−0.2366	−0.1733
x^{ew}	−0.2368	−0.2368	−0.2368	−0.2368
$x^{Li\ et\ al.}$	−0.0713	−0.0243	−0.0505	−0.0936
x^{bl}	−0.0431	−0.1741	−0.1031	−0.2505
CRIX	−0.5552	−0.5552	−0.5552	−0.5552

Table 13
The out-of-sample Sortino ratio of different portfolio strategies ($\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	−0.0848	−0.0981	−0.0680	−0.1418
$x^{Sentiment-ew}$	−0.0611	−0.0611	−0.0611	−0.0611
x^{mv}	−0.1291	−0.1348	−0.1118	−0.1885
$x^{\max}$	−0.2973	−0.3305	−0.3583	−0.2682
x^{ew}	−0.5596	−0.5596	−0.5596	−0.5596
$x^{Li\ et\ al.}$	−0.1078	−0.0376	−0.0763	−0.1402
x^{bl}	−0.0744	−0.2773	−0.1608	−0.3715
CRIX	−0.7458	−0.7458	−0.7458	−0.7458

From the above results, the proposed portfolio strategy $x^{Sentiment}$ has a better out-of-sample performance compared to other portfolio strategies (i.e.,x^{mv},$x^{\max}$,x^{ew},$x^{Li\ et\ al.}$, x^{bl}) and CRIX, and, in most cases., outperforms the strategy $x^{Sentiment-ew}$. Moreover, it proves that forecasting the future movement of cryptocurrency prices is essential for constructing a cryptocurrency portfolio, which can effectively im-prove its out-of-sample performance.

3.3. Robustness test

To test whether the research results in Section 3.2 are robust for different data intervals, we use the test data from July 1, 2021, to August 31, 2021, to test the robustness of the proposed approach (meanwhile, the cryptocurrency market in the data interval selected in Section 3.2 is in a bear market, while in section 3.3 it is in a bull market).

First, we collect daily historical trading data and Google Trends index data for cryptocurrencies (Bitcoin, Ethereum, Ripple, Cardano, Dogecoin, Polkadot, and Litecoin) from January 1, 2021 to August 31, 2021, and collect 32,470,723 tweets. Moreover, we use the multi-source data as SVM input and use the closing price change direction as the output data, then normalize the data and arrange them in chronological order. Next, we use the first six months of data as the training set of the model, and the last two months of data as the test set of the model, with the same tuning process as in Section 3.1. Finally, we obtain the next-day price forecasts, the three-day price forecasts, and the five-day price forecasts for the seven cryptocurrencies.

Similar to Section 3.2, the out-of-sample test is conducted for seven different portfolio strategies and CRIX. We then obtain 62, 60, and 58 out-of-sample portfolio return series by rolling the out-of-sample tests. We can calculate the out-of-sample Sharpe ratio, Sortino ratio, CEQ return, and return-loss for each portfolio strategy across holding periods, then analyze and compare the out-of-sample performance of different portfolio strategies. As in Section 3.2, the “rolling sample” window length is set to 30, 60, 90, and 120, respectively. The results are shown in Tables 15–23.

When the holding period is one day, and the window length M is 30, 60, 90, and 120, respectively, the corresponding results are shown in Table 15 -Table 17. Table 15 and Table 16 show that the proposed portfolio strategy achieves higher Sharpe and Sortino ratios than other portfolio strategies. In addition, in almost all cases, strategies $x^{Sentiment}$ and $x^{Sentiment-ew}$ always have a higher out-of-sample Sharpe ratio Sortino ratio than traditional strategies x^{mv} and x^{ew}. This is consistent with the results in Section 3.2. Meanwhile, even though both incorporate forecasting information, the minimum-variance-based strategy $x^{Sentiment}$ outperforms the strategy $x^{Sentiment-ew}$ in most cases. Moreover, the results of CEQ return show the proposed strategy has a higher CEQ return than traditional portfolio strategies. The results in Panel B show that regardless of the window length, the return-loss of other strategies

Table 14
The out-of-sample CEQ return and return-loss of different portfolio strategies ($\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	−0.0087	−0.0095	−0.0077	−0.0118
$x^{Sentiment-ew}$	−0.0110	−0.0110	−0.0110	−0.0110
x^{mv}	−0.0121	−0.0126	−0.0112	−0.0155
$x^{\max}$	−0.0235	−0.0313	−0.0429	−0.0324
x^{ew}	−0.0358	−0.0358	−0.0358	−0.0358
$x^{Li\ et\ al.}$	−0.0134	−0.0075	−0.0107	−0.0153
x^{bl}	−0.0076	−0.0230	−0.0198	−0.0330
CRIX	−0.0746	−0.0746	−0.0746	−0.0746
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	−0.0019	−0.0029	−0.0005	−0.0066
x^{mv}	0.0029	0.0024	0.0028	0.0029
$x^{\max}$	0.0029	0.0024	0.0028	0.0029
x^{ew}	0.0121	0.0158	0.0264	0.0095
$x^{Li\ et\ al.}$	0.0012	−0.0046	0.0002	−0.0010
x^{bl}	−0.0015	0.0107	0.0066	0.0160
CRIX	0.0213	0.0201	0.0227	0.0161

Table 15
The out-of-sample Sharpe ratio of different portfolio strategies (robustness test, $\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.2610	0.2569	0.2322	0.2407
$x^{Sentiment-ew}$	0.2024	0.2024	0.2024	0.2024
x^{mv}	0.2313	0.2169	0.1731	0.1640
$x^{\max}$	0.1727	0.1901	0.1641	0.1393
x^{ew}	0.1868	0.1869	0.1868	0.1868
$x^{Li\ et\ al.}$	0.0779	0.0320	0.0524	0.0174
x^{bl}	0.2145	0.0774	−0.1129	0.0809
CRIX	−0.0693	−0.0693	−0.0693	−0.0693

Table 16
The out-of-sample Sortino ratio of different portfolio strategies (robustness test, $\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.3858	0.3822	0.3380	0.3592
$x^{Sentiment-ew}$	0.2883	0.2883	0.2883	0.2883
x^{mv}	0.3481	0.3282	0.2655	0.2528
$x^{\max}$	0.2682	0.2966	0.2787	0.2468
x^{ew}	0.3572	0.3572	0.3572	0.3572
$x^{Li\ et\ al.}$	0.1084	0.0460	0.0723	0.0245
x^{bl}	0.4267	0.1157	−0.1200	0.1273
CRIX	−0.0924	−0.0924	−0.0924	−0.0924

Table 17
The out-of-sample CEQ return and return-loss of different portfolio strategies (robustness test, $\tau = 1$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	0.0095	0.0088	0.0073	0.0074
$x^{Sentiment-ew}$	0.0076	0.0076	0.0076	0.0076
x^{mv}	0.0074	0.0064	0.0051	0.0048
$x^{\max}$	0.0065	0.0073	0.0073	0.0064
x^{ew}	0.0069	0.0069	0.0069	0.0069
$x^{Li\ et\ al.}$	0.0022	0.0005	0.0012	0.0000
x^{bl}	0.0178	0.0009	−0.0728	0.0033
CRIX	−0.0057	−0.0057	−0.0057	−0.0057
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	0.0025	0.0023	0.0012	0.0016
x^{mv}	0.0010	0.0013	0.0019	0.0025
$x^{\max}$	0.0038	0.0029	0.0036	0.0059
x^{ew}	0.0031	0.0029	0.0019	0.0022
$x^{Li\ et\ al.}$	0.0067	0.0081	0.0081	0.0083
x^{bl}	0.0052	0.0255	0.0984	0.0124
CRIX	0.0192	0.0189	0.0175	0.0180

Table 18
The out-of-sample Sharpe ratio of different portfolio strategies (robustness test, $\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.4883	0.4560	0.3887	0.3799
$x^{Sentiment-ew}$	0.4298	0.4298	0.4298	0.4298
x^{mv}	0.4419	0.4617	0.3637	0.3511
$x^{\max}$	0.4744	0.4172	0.3165	0.3208
x^{ew}	0.3495	0.3495	0.3495	0.3495
$x^{Li\ et\ al.}$	0.3736	0.3564	0.3670	0.4566
x^{bl}	−0.0602	0.1327	0.1048	0.2763
CRIX	−0.1339	−0.1339	−0.1339	−0.1339

Table 19
The out-of-sample Sortino ratio of different portfolio strategies (robustness test, $\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.7546	0.6976	0.5964	0.5850
$x^{Sentiment-ew}$	0.6396	0.6396	0.6396	0.6396
x^{mv}	0.6852	0.6987	0.5590	0.5405
$x^{\max}$	0.7088	0.6376	0.5112	0.5069
x^{ew}	0.6087	0.6087	0.6087	0.6087
$x^{Li\ et\ al.}$	0.5533	0.5948	0.6066	0.7309
x^{bl}	−0.0646	0.6302	0.1161	0.3882
CRIX	−0.1748	−0.1748	−0.1748	−0.1748

relative to strategy $x^{Sentiment}$ is always positive, which means that the proposed strategy $x^{Sentiment}$ achieves a higher Sharpe ratio than other strategies. It is interesting to see that strategy 1 achieves the highest Sharpe ratio at $M = 30$, but its results are not satisfactory in other cases. Overall, the conclusions in Tables 15–17 are generally consistent with those in Tables 6–8. The proposed strategy remains robust for different data intervals when $\tau = 1$.

When the holding period $\tau = 3$, the results are shown in Tables 18–20. As in the previous results, the proposed strategy $x^{Sentiment}$ achieves the best out-of-sample performance in most cases and only slightly underperforms $x^{Sentiment-ew}$ when the window length is high. Meanwhile, strategy $x^{Sentiment-ew}$ almost consistently outperforms strategy x^{ew}. This suggests that cryptocurrency forecasting information can improve the out-of-sample performance of portfolio strategies effectively. From Table 20, strategy $x^{\max}$ has the highest CEQ return in most cases when $\tau = 3$, but the proposed portfolio strategies still outperform the other strategies. In addition, from Panel B, most of the return-loss of other strategies relative to $x^{Sentiment}$ is positive, which means that $x^{Sentiment}$ achieves the highest Sharpe ratio. This further confirms the previous conclusions. The results in Tables 18–20 show that the proposed strategy $x^{Sentiment}$ still outperforms strategy $x^{Sentiment-ew}$,x^{mv},$x^{\max}$,x^{ew},$x^{Li\ et\ al.}$,x^{bl}, and CRIX in terms of Sharpe ratio, Sortino ratio, etc. in most cases when $\tau = 3$. It is shown that the proposed portfolio strategy remains robust for different data intervals when $\tau = 3$.

Similarly, when the holding period $\tau = 5$, The results are shown in Tables 21–23. Except for strategy x^{mv}, which achieves the highest out-of-sample Sharpe and Sortino ratios at $M = 60$ and $M = 90$ $M = 90$, the proposed strategy achieves a higher out-of-sample Sharpe and Sortino ratios than traditional strategies. This is consistent with the results in Section 3.2 when $\tau = 5$. The return-loss of strategy $x^{Sentiment-ew}$ relative

Table 20
The out-of-sample CEQ return and return-loss of different portfolio strategies (robustness test, $\tau = 3$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	0.0273	0.0247	0.0201	0.0200
$x^{Sentiment-ew}$	0.0258	0.0258	0.0258	0.0258
x^{mv}	0.0259	0.0248	0.0191	0.0188
$x^{\max}$	0.0309	0.0310	0.0261	0.0259
x^{ew}	0.0216	0.0216	0.0216	0.0216
$x^{Li\ et\ al.}$	0.0240	0.0253	0.0261	0.0321
x^{bl}	−0.0832	−4.9210	−0.0052	0.0326
CRIX	−0.0167	−0.0167	−0.0167	−0.0167
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	0.0038	0.0017	−0.0027	−0.0032
x^{mv}	0.0029	−0.0003	0.0014	0.0017
$x^{\max}$	0.0010	0.0032	0.0070	0.0056
x^{ew}	0.0095	0.0073	0.0027	0.0021
$x^{Li\ et\ al.}$	0.0082	0.0080	0.0017	−0.0059
x^{bl}	0.1931	1.0582	0.0712	0.0177
CRIX	0.0577	0.0547	0.0485	0.0477

Table 21
The out-of-sample Sharpe ratio of different portfolio strategies (robustness test, $\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.5408	0.5206	0.4646	0.4824
$x^{Sentiment-ew}$	0.5253	0.5253	0.5253	0.5253
x^{mv}	0.5400	0.5394	0.4808	0.4963
$x^{\max}$	0.3713	0.4004	0.3739	0.4018
x^{ew}	0.4427	0.4427	0.4427	0.4427
$x^{Li\ et\ al.}$	0.3685	0.2854	0.4313	0.5126
x^{bl}	0.0733	0.4219	0.3918	0.4294
CRIX	−0.1952	−0.1952	−0.1952	−0.1952

Table 22
The out-of-sample Sortino ratio of different portfolio strategies (robustness test, $\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
$x^{Sentiment}$	0.8394	0.8275	0.7436	0.7682
$x^{Sentiment-ew}$	0.7780	0.7780	0.7780	0.7780
x^{mv}	0.8512	0.8657	0.7688	0.7895
$x^{\max}$	0.5591	0.6135	0.6218	0.6234
x^{ew}	0.7474	0.7474	0.7474	0.7474
$x^{Li\ et\ al.}$	0.5750	0.5026	0.7487	0.8468
x^{bl}	0.1400	0.7936	0.6620	0.8129
CRIX	−0.2607	−0.2607	−0.2607	−0.2607

to strategy $x^{Sentiment}$ is negative when the window length is large, which is consistent with that in Section 3.2 when $\tau = 5$. It is shown that when $\tau = 5$, the strategy $x^{Sentiment-ew}$ has better out-of-sample performance than the strategy $x^{Sentiment}$. Overall, the results of Tables 21–23 show that when the holding period is 5 days, the proposed strategy still outperforms other strategies and CRIX in terms of Sharpe ratio, Sortino ratio, etc. in most cases.

From the results in Tables 15–23, the out-of-sample performance of strategy $x^{Sentiment}$,$x^{Sentiment-ew}$,x^{mv},$x^{\max}$, and x^{ew} tends to decrease as the window length M increases, which further confirms the findings in Section 3.2. It suggests that investors should not rely too heavily on the historical performance of assets when constructing a cryptocurrency portfolio. And as for strategies $x^{Li\ et\ al.}$ and x^{bl}, their out-of-sample performance is inconsistent, indicating that these strategies may not suitable for the highly volatile cryptocurrency market. In addition, the

Table 23
The out-of-sample CEQ return and return-loss of different portfolio strategies (robustness test, $\tau = 5$).

Rolling sample	$M = 30$	$M = 60$	$M = 90$	$M = 120$
Panel A: Certainty-equivalent return of portfolio returns				
$x^{Sentiment}$	0.0419	0.0370	0.0328	0.0344
$x^{Sentiment-ew}$	0.0455	0.0455	0.0455	0.0455
x^{mv}	0.0425	0.0400	0.0346	0.0359
$x^{\max}$	0.0276	0.0323	0.0431	0.0401
x^{ew}	0.0372	0.0372	0.0372	0.0372
$x^{Li\ et\ al.}$	0.0317	0.0255	0.0452	0.0501
x^{bl}	−0.2807	0.0623	0.0454	0.0537
CRIX	−0.0311	−0.0311	−0.0311	−0.0311
Panel B: Return-loss relative to $x^{Sentiment}$				
$x^{Sentiment-ew}$	0.0014	−0.0005	−0.0058	−0.0041
x^{mv}	0.0001	−0.0015	−0.0013	−0.0011
$x^{\max}$	0.0142	0.0109	0.0129	0.0094
x^{ew}	0.0092	0.0073	0.0021	0.0037
$x^{Li\ et\ al.}$	0.0171	0.0261	0.0041	−0.0033
x^{bl}	0.3862	0.0188	0.0103	0.0081
CRIX	0.0893	0.0869	0.0801	0.0822

above results also show that the proposed portfolio strategy $x^{Sentiment}$ can effectively improve the out-of-sample performance of portfolios under different holding periods. Moreover, the proposed portfolio optimization model remains robust for different data intervals, which further proves the effectiveness and practicality of this approach.

3.4. Out-of-sample performance with transaction costs

This section discusses the impact of transaction costs on each portfolio strategy. Since we found that most of the strategies achieved the best results when $M = 30$ in our previous study, we chose the Sharpe ratio results for each strategy at $M = 30$ as the reference basis, we include transaction costs in the portfolio and perform an out-of-sample test to obtain the corresponding Sharpe ratio, and we set the transaction costs to 0.1 % concerning the real cryptocurrency market. The results are shown below:

As can be seen from the results in Figs. 4 and 5, even with transaction costs, the proposed strategy still has the highest out-of-sample performance in most cases, which further illustrates that the proposed cryptocurrency portfolio strategy can guide investing in the real cryptocurrency market.

4. Conclusion

By combining portfolio theory, text sentiment analysis, and machine learning, this paper proposes a novel cryptocurrency portfolio optimization model, which can improve the out-of-sample performance of cryptocurrency portfolio strategies. First, we employ multiple sources (i.e., historical trading data, Google Trends, and tweets from Twitter) to predict the future movement of cryptocurrency prices. To this end, we crawl all tweets related to cryptocurrencies from Twitter within a certain period and apply the VADER algorithm to analyze the tweets for the sentiment. Second, we construct three different sentiment indicators, and then employ the multi-source data to forecast the future price movements of cryptocurrencies by using the SVM algorithm. Finally, we propose a novel portfolio optimization model by considering both the forecasting results of cryptocurrencies and the minimum variance portfolio model. Moreover, we provide investors with a corresponding investment strategy. In the empirical part, we select seven representative cryptocurrencies (i.e., Bitcoin, Ethereum, Ripple, Cardano, Dogecoin, Polkadot, and Litecoin) as the underlying assets and then perform price forecasting for these assets. Moreover, we use the forecasting results and historical data of cryptocurrencies for out-of-sample tests with different holding periods, respectively. The

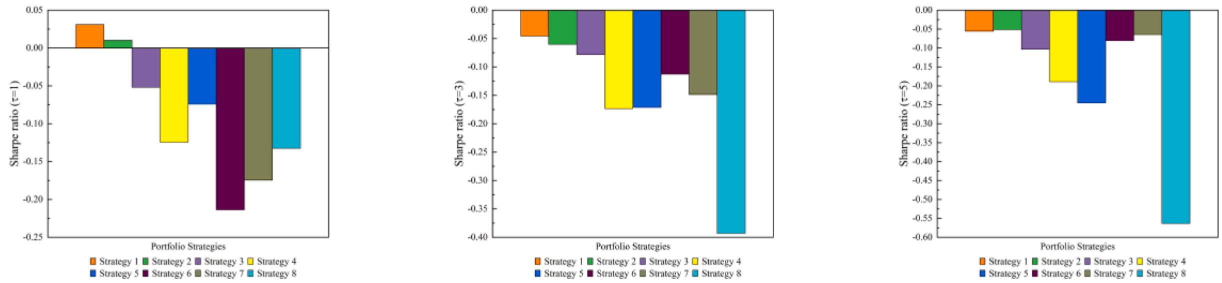

Fig. 4. Sharpe ratio of each strategy with transaction costs ($M = 30$).

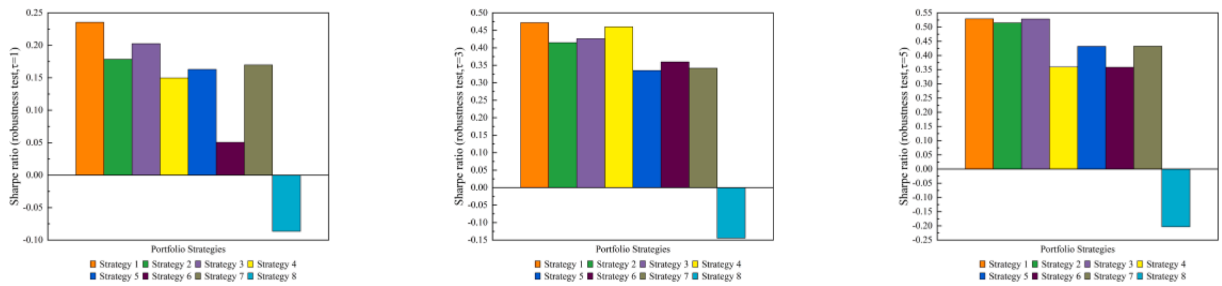

Fig. 5. Sharpe ratio of each strategy with transaction costs (robustness test $M = 30$).

empirical results show that the out-of-sample performance of the proposed portfolio strategy outperforms other portfolio strategies and the CRIX in most cases. Then, we perform a robustness test to validate the above findings. Finally, we discuss the out-of-sample performance with transaction costs. Furthermore, this paper provides several practical suggestions for investors in constructing their portfolios.

On the one hand, when constructing a cryptocurrency portfolio strategy, we should not rely too heavily on the historical characteristics of the assets, as they may not be consistent with future characteristics. The future price forecasting information of cryptocurrencies obtained by using multi-source data can reflect the future movements of cryptocurrencies to a certain extent. It can be incorporated into portfolio optimization and improve the effectiveness and practicality of portfolio strategy.

On the other hand, the out-of-sample Sharpe ratio, Sortino ratio, and CEQ return for most strategies decrease as the window length increases. It suggests that one does not need to consider long-term historical information when constructing a cryptocurrency portfolio, but should focus on the short-term historical performance of assets, which may be due to the high volatility of cryptocurrency assets themselves.

This study still has some limitations, and the current work in this paper could be further extended from the following two aspects (i) We only focus on cryptocurrency price movements; however, it does not reflect the magnitude of the price movements; therefore, subsequent studies can include the analysis of the magnitude of the price movements. (ii) Only seven mainstream cryptocurrencies are used for the portfolio and can consider adding more cryptocurrency assets in the future and constructing corresponding asset evaluation methods and selection methods to select high-quality cryptocurrency assets for the portfolio.

CRediT authorship contribution statement

Zhongbao Zhou: Conceptualization, Writing – review & editing, Supervision. **Zhengyang Song:** Conceptualization, Writing – original draft, Writing – review & editing. **Helu Xiao:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Tiantian Ren:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the Hunan Key Laboratory (grant number. 2020TP1013), Hunan Provincial Social Science Foundation (grant number. 20YBA060) and Key Project of Scientific Research of Hunan Provincial Education Department (grant number. 21A0020).

References

Adcock, R., & Gradojevic, N. (2019). Non-fundamental, non-parametric Bitcoin forecasting. *Physica A: Statistical Mechanics and its Applications, 531*, Article 121727.
Aharon, D. Y., Demir, E., Lau, C. K. M., & Zaremba, A. (2022). Twitter-Based uncertainty and cryptocurrency returns. *Research in International Business and Finance, 59*, Article 101546.
Aljinović, Z., Marasović, B., & Šestanović, T. (2021). Cryptocurrency Portfolio Selection—A Multicriteria Approach. *Mathematics, 9*(14), 1677.
Almansour, B. Y., Alshater, M. M., & Almansour, A. Y. (2021). Performance of ARCH and GARCH Models in Forecasting Cryptocurrency Market Volatility. *Industrial Engineering & Management Systems An International Journal, 20*(2), 130–139.
Andrei, D., & Hasler, M. (2015). Investor Attention and Stock Market Volatility. *The Review of Financial Studies, 28*(1), 33–72.
Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance, 59*(3), 1259–1294.
Arias, M., Arratia, A., & Xuriguera, R. (2013). Forecasting with Twitter Data. *ACM Transactions on Intelligent Systems and Technology, 5*(1), 1–24.
Baek, C., & Elbeck, M. (2015). Bitcoins as an investment or speculative vehicle? *A first look. Applied Economics Letters, 22*(1), 30–34.
Bijl, L., Kringhaug, G., Molnar, P., & Sandvik, E. (2016). Google searches and stock returns. *International Review of Financial Analysis, 45*, 150–156.
Bouri, E., Gupta, R., Tiwari, A. K., & Roubaud, D. (2017). Does Bitcoin hedge global uncertainty? Evidence from wavelet-based quantile-in-quantile regressions. *Finance Research Letters, 23*, 87–95.
Bouri, E., Lucey, B., & Roubaud, D. (2020). Cryptocurrencies and the downside risk in equity investments. *Finance Research Letters, 33*, Article 101211.

Bouri, E., Molnár, P., Azzi, G., Roubaud, D., & Hagfors, L. I. (2017). On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier? *Finance Research Letters, 20*, 192–198.

Brauneis, A., & Mestel, R. (2018). Price discovery of cryptocurrencies: Bitcoin and beyond. *Economics Letters, 165*, 58–61.

Brauneis, A., & Mestel, R. (2019). Cryptocurrency-portfolios in a mean-variance framework. *Finance Research Letters, 28*, 259–264.

Brière, M., Oosterlinck, K., & Szafarz, A. (2015). Virtual currency, tangible return: Portfolio diversification with bitcoin. *Journal of Asset Management, 16*(6), 365–373.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. (2020). Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. *Journal of Medical Internet Research, 22*(10), 22624.

Chopra, V. K., & Ziemba, W. T. (1993). The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice. *The Journal of Portfolio Management, 19*(2), 6.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Deb, S. (2021). Analyzing airlines stock price volatility during COVID-19 pandemic through internet search data. *International Journal of Finance & Economics.*

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies, 22*(5), 1915–1953.

Dey, A. K., Hoque, G. M. T., Das, K. P., & Panovska, I. (2022). Impacts of COVID-19 local spread and Google search trend on the US stock market. *Physica A, 589*, 126423-126423.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. *Applied Sciences-Basel, 10*(21), 7748.

Giudici, P., Pagnottoni, P., & Polinesi, G. (2020). Network Models to Enhance Automated Cryptocurrency Portfolio Management. *Frontiers in artificial intelligence, 3*, 22-22.

Glaser, F., Zimmermann, K., Haferkorn, M., & Weber, M. C. (2014). Bitcoin-asset or currency? revealing users' hidden intentions. *Proceedings of the ECIS 2014 (Tel Aviv)*.

Guesmi, K., Saadi, S., Abid, I., & Ftiti, Z. (2019). Portfolio diversification with virtual currency: Evidence from bitcoin. *International Review of Financial Analysis, 63*, 431–437.

Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior & Organization, 117*, 62–81.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research, 32*(10), 2513–2522.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225.

Ibrahim, A. (2021). Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model. *Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*.

Jagannathan, R., & Ma, T. (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance, 58*(4), 1651–1683.

Kajtazi, A., & Moro, A. (2019). The role of bitcoin in well diversified portfolios: A comparative global study. *International Review of Financial Analysis, 61*, 143–157.

Klein, T., Hien Pham, T., & Walther, T. (2018). Bitcoin is not the New Gold - A comparison of volatility, correlation, and portfolio performance. *International Review of Financial Analysis, 59*, 105–116.

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money, 65*, Article 101188.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance, 10*(5), 603–621.

Li, Y., Bu, H., Li, J., & Wu, J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting, 36*(4), 1541–1562.

Liang, C., Zhang, Y., Li, X., & Ma, F. (2020). Which predictor is more predictive for Bitcoin volatility? And why? *International Journal of Finance & Economics*, 1947–1961.

Liu, W. (2019). Portfolio diversification across cryptocurrencies. *Finance Research Letters, 29*, 200–205.

Liu, Y. K., & Tsyvinski, A. (2021). Risks and Returns of Cryptocurrency. *Review of Financial Studies, 34*(6), 2689–2727.

Ma, Y., Ji, Q., & Pan, J. (2019). Oil financialization and volatility forecast: Evidence from multidimensional predictors. *Journal of Forecasting, 38*(6), 564–581.

Mahdi, E., Leiva, V., Mara'Beh, S., & Martin-Barreiro, C. (2021). A New Approach to Predicting Cryptocurrency Returns Based on the Gold Prices with Support Vector Machines during the COVID-19 Pandemic Using Sensor-Related Data. *Sensors, 21* (18), 6319.

Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics, 135*(2), 293–319.

Mehlawat, M. K., Gupta, P., & Khaitan, A. (2021). Multiobjective fuzzy vehicle routing using Twitter data: Reimagining the delivery of essential goods. *International Journal of Intelligent Systems, 36*(7), 3566–3595.

Michaud, R. O. (1989). The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Financial Analysts Journal, 45*(1), 31–42.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review, 21260*.

Naseer, A., Nava Baro, E., Daud Khan, S., Vila, Y., & Doyle, J. (2022). A Novel Cryptocurrency Prediction Method Using Optimum CNN. *Computers, Materials & Continua, 71*(1), 1051–1063.

Niu, F., Nie, S., & Wang, W. (2008). The Forecasts Performance of Gray Theory, BP Network, SVM for Stock Index. *Proceedings of the 2008 International Symposium on Knowledge Acquisition and Modeling.*

Nofer, M., & Hinz, O. (2015). Using Twitter to Predict the Stock Market Where is the Mood Effect? *Business & Information Systems Engineering, 57*(4), 229–242.

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications, 73*, 125–144.

Qadan, M., & Nama, H. (2018). Investor sentiment and the price of oil. *Energy Economics, 69*, 42–58.

Sahamkhadam, M., Stephan, A., & Östermark, R. (2022). Copula-based Black-Litterman portfolio optimization. *European Journal of Operational Research, 297*(3), 1055–1070.

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters, 32*, Article 101084.

Trimborn, S., & Härdle, W. K. (2018). CRIX an Index for cryptocurrencies. *Journal of Empirical Finance, 49*, 107–122.

Trimborn, S., Li, M., & Härdle, W. K. (2020). Investing with Cryptocurrencies—a Liquidity Constrained Investment Approach*. *Journal of Financial Econometrics, 18* (2), 280–306.

Tsukioka, Y., Yanagi, J., & Takada, T. (2018). Investor sentiment extracted from internet stock message boards and IPO puzzles. *International Review of Economics & Finance, 56*, 205–217.

Tu, J., & Zhou, G. (2011). Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics, 99*(1), 204–215.

Ustun, O., & Kasimbeyli, R. (2012). Combined forecasts in portfolio optimization: A generalized approach. *Computers & Operations Research, 39*(4), 805–819.

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance, 36*(6), 1808–1821.

Wang, H. Y., Li, H., & Shen, J. Y. (2019). A Novel Hybrid Fractal Interpolation-SVM Model for Forecasting Stock Price Indexes. *Fractals-Complex Geometry Patterns and Scaling in Nature and Society, 27*(4), 1950055.

Wang, W., Li, W., Zhang, N., & Liu, K. (2020). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications, 143*, Article 113042.

Wu, C. (2021). Window effect with Markov-switching GARCH model in cryptocurrency market. *Chaos Solitons & Fractals, 146*, Article 110902.

Yun, C., & s., & Shin, J.. (2020). Forecasting Cryptocurrency Volatility Using a MS-EGARCH Model. *Korean Journal of Financial Engineering, 19*(2), 1–22.

Zhang, Z., Dai, H., Zhou, J., Mondal, S. K., García, M. M., & Wang, H. (2021). Forecasting cryptocurrency price using convolutional neural networks with weighted and attentive memory channels. *Expert Systems with Applications, 183*, Article 115378.

Zhou, Z., Gao, M., Xiao, H., Wang, R., & Liu, W. (2021). Big data and portfolio optimization: A novel approach integrating DEA with multiple data sources. *Omega, 104*, Article 102479.