International Journal of Computational Intelligence Systems (2024) 17:73
https://doi.org/10.1007/s44196-024-00465-0

RESEARCH ARTICLE

Semiconductor Price Index Predicting Based on a Novel Improved AdaBoost Feature-Weighted Combination Model

Feng Chen[1] · Qi Jiang[1] · Hongyu Deng[1]

Received: 11 January 2024 / Accepted: 20 March 2024

Abstract
The semiconductor price index serves as a vital metric for assessing technological developments and related market trends. Establishing a more accurate forecasting model for the semiconductor price index is of significant importance for analyzing the industry's trends and market directions. In this paper, a novel framework for semiconductor price index forecasting is proposed. In addition to traditional financial data, the study introduces search engine data (Google Trends) representing investor attention, and introduces text information extracted from online news headlines reflecting major market events and government policies as independent variables. Used to predict the dependent variable: The PHLX Semiconductor Sector (SOX). First, the XGBoost model is employed to compute the importance scores of each feature. Then, a feature weight coefficient indicator is constructed based on these importance scores to calculate the weight coefficient indicator values for each feature. These indicator values are then used to weight the kernel function of Support Vector Regression (SVR), resulting in weighted Support Vector Regression (WSVR). Finally, WSVR is utilized as the base learner for Adaptive Boosting (AdaBoost), yielding the XGBoost–WSVR–AdaBoost model based on feature weighting. The proposed model outperforms AdaBoost, RNN, ERT, LSTM, and other models in terms of Mean Absolute Percentage Error (MAPE) and goodness-of-fit (R^2). It also exhibits superior predictive performance compared to models in ablation experiments, and the introduction of text data or Google trends further improves the prediction performance of the model. In conclusion, the improved AdaBoost feature-weighted combination model proposed in this study offers a more accurate prediction for semiconductor price index.

Keywords Semiconductor price index · XGBoost–WSVR–AdaBoost · Google trends · Text analysis

1 Introduction

As the "heart" of electronic information products, and the crown jewel of the modern industrial technology system, semiconductors are increasingly used in all aspects of the national economy and social life, guiding artificial intelligence, big data, autonomous driving, etc. The rise of new industries is the underlying support of the digital age. It is of great strategic significance to a country's economic growth, technological progress, national defense and security, upgrading of industrial structure and improvement of core competitiveness.

The semiconductor industry was born in the United States in 1958, with IBM becoming the first company to mass-produce transistors. The semiconductor industry has a large investment scale, fast update speed (following the "Moore effect"), high technology intensity, high intensity of R&D activities, long cycles, and long innovation chains. At the same time, the semiconductor industry has high industry barriers and is a typical capital-technology dual-intensive industry [1]. On the one hand, new chip demand markets such as autonomous vehicles, artificial intelligence, and the industrial Internet are constantly being created by digitalization. On the other hand, smart homes, virtual reality, etc. are also increasingly popular in the digital era, leading to rising global chip demand. According to IC Insights, the global chip market is expected to exceed US$1 trillion by 2030. The huge gap in chip demand has also put forward structural adjustment requirements for the global semiconductor supply chain.

In recent years, the development and growth of the semiconductor industry has attracted widespread attention and investment. Semiconductor industry stocks

✉ Feng Chen
chenfengmath@163.com

[1] School of Mathematics and Statistics, Changchun University, Changchun, China

Published online: 02 April 2024

are performing well. Investors have great enthusiasm for investing in the semiconductor field. The prediction of semiconductor stock prices is valuable and meaningful.

In the field of numerical prediction, the scope and methods of machine learning applications are constantly evolving. For example, in epidemiological prediction, Tutsoy et al. [2] proposed a promising solution for large-scale epidemic prediction using graph theory and metaheuristics. Chen et al. [3] applied similar techniques in the field of traffic data interpolation and established the NT-DPTC model. Domestic and foreign scholars have used different methods to study the semiconductor industry, such as prediction of semiconductor-related prices based on machine learning algorithms [4, 5], prediction of semiconductor industry stocks based on neural networks [6, 7], prediction of semiconductor stocks based on network big data [8]. To predict semiconductor stock prices more accurately, this article proposes a new prediction framework.

1.1 Prediction of Semiconductor Price

Semiconductor price forecasting is receiving increasing attention from the academic community. Semiconductor price forecasting entails complexity due to factors such as insufficient capacity, cost constraints, and the intricacies of semiconductor manufacturing processes. Artificial Intelligence (AI) technology holds greater applicability in the field of semiconductor price forecasting [9, 10].

Predicting semiconductor stock prices falls within the realm of financial time series forecasting, which has long been a focal point of research and finds widespread application [11]. Traditional models like AR, MA, ARMA, and ARIMA have been widely utilized * MERGEFORMAT [4, 12, 13]. However, these models often assume data follows specific distributions and struggle to handle issues, such as nonlinearity, non-stationarity, and heteroscedasticity effectively. In addition, these models are based solely on the characteristics of the data and may not capture the influence of other factors [12].

With the continuous development of machine learning technology, commonly used models such as neural networks, support vector machines, and random forests can better handle nonlinear, non-stationary, and heterogeneous data. Convolutional neural networks, recurrent neural networks, and long–short-term memory networks possess powerful feature extraction and representation capabilities, and demonstrate good predictive performance in certain metrics. The development of machine learning has provided additional solutions for prediction. Chen [5] utilized fuzzy theory to establish a multivariate linear regression model for forecasting DRAM product prices. Chen and Wu [14] employed the Partial Consensus Fuzzy Intersection (PCFI) operator instead of the Fuzzy Intersection (FI) operator to aggregate expert fuzzy forecasts, and applied it to predict the unit cost of DRAM products. Fuzzy theory and deep learning have also been cited in practical applications of semiconductor price forecasting [5, 6, 15]. Li et al. [16] utilized long−short-term memory (LSTM) for financial data forecasting. Xu [6] predicted the rise and fall of global semiconductor industry stocks using artificial neural networks (ANN). All three models performed well, with ANN showing the best results. To improve model effectiveness, thorough comparisons of deep learning methods are necessary.

Combination models integrate multiple machine learning algorithms to form a stronger model, aiming to improve the accuracy and reliability of predictions and decisions [8]. Increasingly, scholars are applying combination models to forecasting tasks [17]. Scholars have explored the fusion of traditional time series methodologies with machine learning approaches [16], as well as the amalgamation of various machine learning techniques [9, 18]. In addition, researchers have investigated the integration of machine learning methodologies with genetic algorithms [19]. Xia et al. [20] introduced a novel approach utilizing an improved GRU–RNN for predicting renewable energy generation and power load under both univariate and multivariate scenarios. Wang et al. [21] developed a hybrid model using ARIMA and XGBoost for forecasting domain model improvement. Grau et al. [7] employed long–short-term cognitive networks (LSTCN) to forecast demand for six different types of semiconductor company products. The results indicated that LSTCN exhibited superior predictive performance and predicted the peaks and troughs of demand more accurately. While considering economic indicators, incorporating search engine data has shown promising results. However, there is still significant potential for improvement in forecasting accuracy.

1.2 Text Technology and Google Trends Research

Sentiment analysis of news texts is to analyze and mine the emotional information contained in news texts, to provide more accurate information services for news media, and to provide strong support for people to understand social public opinion. Introducing interference factor modeling can weigh the uncertainty of data [22], and news text can effectively reflect uncertainty. Currently, more and more scholars are applying news text analysis to various fields.

In terms of variable processing, the most traditional text sentiment method classifies editorial articles as “bullish”, “bearish”, and “uncertain” and uses categories as variables for prediction [23]. With the continuous development of text analysis technology, there are more abundant methods for text emotion processing. Li et al. [24] built a dictionary of positive and negative emotions, a dictionary of negative

words, a dictionary of degree adverbs, and a dictionary of transitions. Cut the news text according to the rules, and divided the news into meaning groups, sentences, paragraphs, and chapters. Through the formulated rules Calculate the emotional tendency value of key sentences, and finally obtain the emotional tendency value of the paragraph and the entire article, thereby obtaining the emotional tendency of the news. Wu et al. [25] proposed an emerging industry news text monitoring method that combines the structured topic model (STM) with covariates and deep learning sentiment analysis technology. By monitoring the changes in the intensity of industry news hot spots reported by the media, the emotional tendency of the text has an impact on the news. The temporal impact of hot spot intensity is used to discover and track hot spots and trends in emerging industries. The study found that changes in the media's emotional inclination toward the issuance and trading of various digital tokens, from praise to disparagement, can play an early warning role in the hidden risks of the blockchain.

The continuous innovation of text processing technology has further promoted the development of news text sentiment analysis. Lin et al. [26] proposed a clustering method based on the correlation coefficient of probabilistic language term sets, providing an effective modeling method. Lu et al. [27] proposed a topic change point detection (Topic-CD) model to define topic change points from the perspective of hyperparameters related to topic-word distribution. Rashid et al. [28] proposed a new fuzzy topic modeling (FTM) method to improve the sparsity of short documents. With the development of neural networks, neural network models such as LSTM have more and more applications in this field. Gao et al. [29] built an attention model by improving the TF-IDF algorithm and Bi-LSTM neural network, then used the roBerta pre-training model to well represent the text information, and finally used the LSTM–GRU deep neural network to make polarity classification judgments. It is proved that the performance of this algorithm is better than the traditional algorithm to a certain extent. Xu and Tian [30] applied the BERT model to financial news sentiment analysis, combined emotional characteristics with stock market trading data, and established an LSTM model. They proved that after introducing text sentiment, the model prediction effect was higher than that of three benchmark models (BP neural network, support vector machine, XGBoost). Xu et al. [31] used the BERT pre-training model to perform word embedding mapping on the text, then used the BiLSTM–CNN model to further extract text context and local key features, and finally classified the news text. Zhai et al. [32] combined the advantages of convolutional neural network (CNN) and recurrent neural network (RNN) to propose a hybrid model of CNN and LSTM. Experimental results show that the accuracy of the CNN–LSTM model is 99% in the training set and 92.52% in the validation set. The accuracy of this model is about 3% higher than the LSTM model and about 5% higher than the TextCNN model, indicating that the model proposed in the paper can better express the original features of the data and help improve the effect of text classification.

In recent years, massive network big data has enriched related research on time series. At the same time, search engines are the most important tool for the public to obtain information from the Internet in China, and Google Trends is an indicator representing public concern [33, 34]. In financial analysis, whether it is long-term, medium-term or short-term prediction, the introduction of Google Trends has great potential to improve the effect of prediction models [35, 36]. Google Trends data need to be integrated in practical applications for practical application [37]. Yang et al. [38] proposed a K-means-KPCA–KELM hybrid crude oil price prediction method, combining economic variables and oil-related Google Trends as a series of independent variables. Through numerical experiments, it was found that the accuracy of the model after the introduction of Google Trends on the basis of improving the accuracy of the hybrid model, it can be further improved. Li et al. [39] established an artificial neural network model based on S&P 500 index call options, introduced 25 Internet search attention indicators, integrated them into a Google Trends index, and added the change rate of the Google Trends index to the artificial neural network. The network model improves the estimation accuracy by about 30%. It also revealed that the impact of the Google Trends Index change rate on changes in implied volatility is nonlinear, and the change rate indicates the magnitude of future changes in implied volatility.

1.3 Research Routes

This paper adopts multiple financial indicators related to the semiconductor industry, semiconductor upstream raw materials and semiconductor downstream enterprises, and uses text mining technology to mine the hidden information of semiconductor news headlines to construct the text indicators, and obtains the Google trend indicators representing investor attention. To further improve the prediction performance of the PHLX Semiconductor Sector (SOX), this paper also constructs a combination model based on feature-weighting XGBoost–WSVR–AdaBoost.

This article conducts research from the following three aspects: first, it introduces the construction method of text indicators and the processing method of Google Trends, and explains the design ideas of the XGBoost–WSVR–AdaBoost model. Second, the data sources are described, relevant features are constructed, and the evaluation indicators and parameter selection methods are introduced. Finally, a comparative analysis of the prediction results was conducted

to verify the effectiveness of text indicators, Google Trends indicators, and the effectiveness of the new model.

This article introduces neural network models that have been used frequently in recent years for comparison, and introduces news text indicators and Google trends indicators to enrich the information brought by traditional economic variables, which has sufficient theoretical basis and experimental value. The overall research framework of this article is shown in Fig. 1.

The main contributions of this paper can be summarized as follows:

(1) Based on traditional financial data, this article explores the use of text data and search engine data as foreign aid data to predict the semiconductor price index. It also provides a reference for expanding data sources for other related studies.
(2) This article constructs a feature weight coefficient index, which can effectively weight the feature

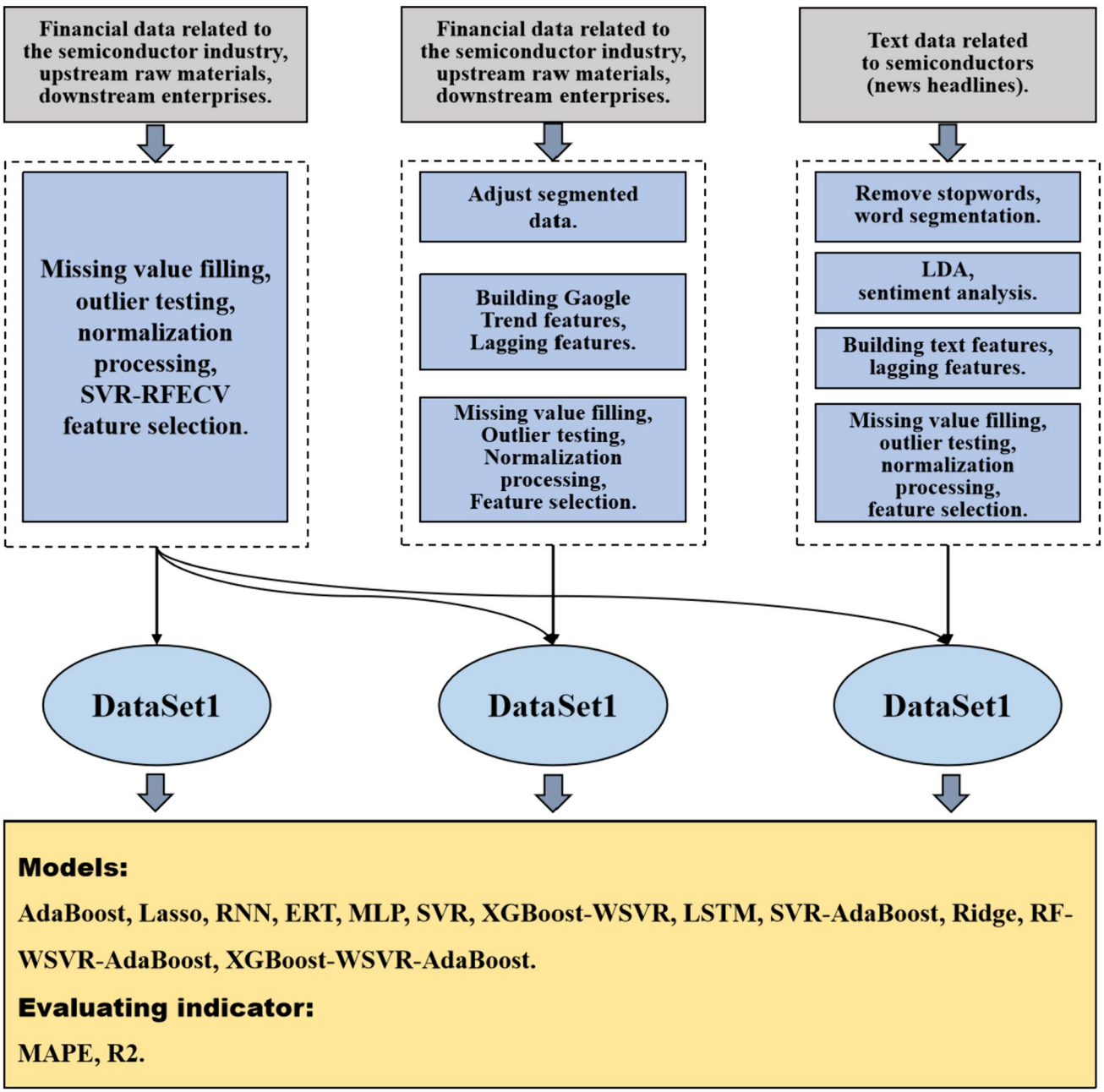

Fig. 1 Overall research framework

variables in the model and further improve the prediction accuracy of the model.

(3) This article proposes a feature weighted combination model. The process of creating the model includes: applying XGBoost to calculate the importance score of each feature variable; calculating the result of each feature variable under the feature weight coefficient index created in this article; weighting the Gaussian kernel function in SVR based on this result; weighting SVR As the benchmark model of AdaBoost, the XGBoost–WSVR–AdaBoost combination model based on feature weighting is finally constructed.

2 Model Proposed

This paper proposes a feature weight index, combining extreme gradient boosting (XGBoost), weighted support vector regression model (WSVR) and adaptive boosting (AdaBoost), and proposes a feature-weighted XGBoost–WSVR–AdaBoost model to improve forecasting performance of the semiconductor price index. On this basis, news text information and Google Trends data are added to further improve the prediction performance. This chapter will introduce the construction method of text indicators, the processing method of Google Trends data, and the design process of XGBoost–WSVR–XGBoost–AdaBoost model.

2.1 Construction of Text Indicators

This article uses the SnonNLP library in Python to calculate the sentiment score of each news headline. News with a sentiment score greater than 0.5 are represented as 1, indicating positive sentiment; values less than 0.5 are represented as 0, indicating negative sentiment; values equal to 0.5 are represented as 2, indicating no obvious emotional tendency. The LDA topic model is used to classify text topics and mine the information contained in different topics.

Dividing each news title according to the time interval of 1 day, we can get the daily news number M_t, the daily positive sentiment news number M_t^{pos}, and the daily negative sentiment news number M_t^{neg}. Average daily sentiment scores across different topics can also be calculated. On the basis of constructing the above text features, this article introduces the features B_t, B_t^* and B_t^{Att} [16, 40] constructed based on the information classification results.

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

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

$$B_t^{\text{Att}} = B_t \ln\left(1 + M_t\right). \tag{3}$$

Among them, M_t^c represents the number of messages of different categories, and M_t represents the total amount of information.

2.2 Processing of Google Trends Data

Google only provides daily data with a time interval of less than 6 months, and the sample time span of this article spans more than 3 years. Therefore, daily Google Trends for the entire time range are not directly available and must be downloaded in segments. The value of Google Trends data reflects the number of searches for a specific term relative to the total number of Google searches, normalized and indexed to a range of 0–100 during the selected time period, where 100 represents the greatest search interest during the selected time period.

If the data are directly spliced after segmented downloading, the trend value cannot fully reflect the search trend of a certain keyword within the entire time range. Therefore, it is necessary to adjust the various Google Trends data after segmented downloads to obtain daily Google Trends data over a reasonable full time period.

Referring to the Google Trends processing method proposed by Xu et al. [37], this paper processes daily frequency Google Trends data in three steps: (1) Download daily Google Trends data for half a year starting from January 1, 2020. (2) Download the daily data for the other half year. Set the overlap between the two periods to 3 months to recover Google Trends more accurately. (3) The score of the ith Google search in the overlapping part of the previous period is overlap_i^{j-1}, and the score of the Google search corresponding to the next period is overlap_i^j. The adjusted Google Trends time series $\widehat{S}_t$ is calculated as follows:

$$\widehat{S}_t = S_t^j \cdot \omega^j, \tag{4}$$

$$\omega^j = \frac{1}{n}\sum_{i=1}^{n} \frac{\text{overlap}_i^{j-1}}{\text{overlap}_i^j}, \tag{5}$$

where S_t^j is the search score (Google Trends) at time t in the j-th cycle, and ω^j is the weight of the j-th cycle.

The adjustment diagram of Google Trends is shown in Fig. 2.

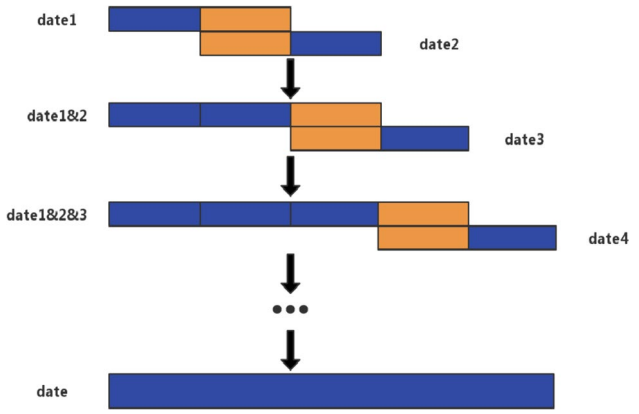

Fig. 2 Google trends adjustment diagram

2.3 Design of XGBoost–WSVR–AdaBoost Model

The traditional AdaBoost algorithm adjusts the weights of samples and base learners, and integrates several weak learners into strong learners after multiple iterations to improve prediction performance. However, this algorithm does not consider the different effects of different features of the model on the prediction results. This paper constructs a new feature weight coefficient index, and on this basis combines extreme gradient boosting (XGBoost), weighted support vector regression model (WSVR) and adaptive boosting (AdaBoost), and proposes a feature weighted XGBoost–WSVR–AdaBoost model. The specific implementation process of this algorithm is as follows:

For the training set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}, x_i \in R^n, y_i \in R$, the number of base learners is T.

(i) Build the XGBoost model and use model.feature_importances_ in Python to obtain the importance score of each feature:

$$u_i (i = 1, 2, \cdots, n). \tag{6}$$

Since the gap in the importance scores of each feature is relatively large, and the feature screening step has been completed before modeling and prediction, the features that need to be weighted here have been considered effective features, and we do not want the gap between the weights of each feature to be too large. Therefore, after using XGBoost to obtain the importance score u_i of each model, further processing is required. This article first sorts the importance score u_i of each feature to obtain the ranking v_i of each feature. Then a feature weight coefficient index ω_i is constructed based on the ranking v_i.

(ii) Sort the feature importance scores u_i to obtain the importance ranking of each feature variable:

$$v_i (i = 1, 2, \cdots, n). \tag{7}$$

(iii) Construct the feature weight coefficient index ω_i:

$$\omega_i = e^{-\frac{v_i}{n}} + 1, i = 1, 2, \cdots, n. \tag{8}$$

(iv) Construct the diagonal matrix P of ω_i:

$$P = \begin{bmatrix} \omega_1 & & \\ & \omega_i & \\ & & \omega_n \end{bmatrix}. \tag{9}$$

(v) Construct an RBF kernel function with feature weighting:

$$\begin{aligned} K_{p-RBF}(x_i, x_j) &= \exp\left(-\frac{1}{2\sigma^2}\|Px_i - Px_j\|^2\right) \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[(x_i - x_j)^T PP^T (x_i - x_j)\right]\right\}. \end{aligned} \tag{10}$$

where the weight matrix P is the diagonal matrix of order n, σ is the RBF kernel function parameter, $x_i \in R^n$ is the ith training sample, and $x_j \in R^n$ is the jth training sample.

Since AdaBoost is an ensemble algorithm based on bagging, when studying the feature weight of the AdaBoost model, only its benchmark model can be weighted. The benchmark model selected in this article is Support Vector Regression (SVR). This is because the support vector machine can transform the huge inner product operation in the high-dimensional space into function evaluation in the low-dimensional space by introducing the kernel function, which simplifies the calculation at the same time. It also avoids the difficulty of finding mapping functions. This article uses the RBF kernel function as the kernel function of support vector regression. Therefore, the feature weighting of AdaBoost is the feature weighting of SVR, which is the weighting of the RBF kernel function.

Finally, after obtaining the feature-weighted RBF kernel function, a weighted support vector regression machine model can be constructed based on this kernel function. That is, WSVR, denoted as ζ, and ζ can be used as the base model of AdaBoost.

(vi) Build an AdaBoost model based on ζ. First, initialize the weight distribution of training samples:

$$D_i = (\omega_{11}, \omega_{12}, \cdots, \omega_{1m}), \omega_{1i} = \frac{1}{m}, i = 1, 2, \cdots, m. \tag{11}$$

where ω_{1i} is the weight of each sample and m is the number of samples.

The following t represents the iteration round. For t, we set $t = 1, 2, \cdots, T$.

(a) Train the base learner using the training data set with the current distribution D_i

$$h_t = \zeta\left(D, D_t\right). \tag{12}$$

(b) Calculate the maximum sample error in the training set:

$$E_t = \max\left|y_i - h_t\left(x_i\right)\right|, i = 1, 2, \cdots, m. \tag{13}$$

where $h_t\left(x_i\right)$ is the predicted value of the weak learner for the i-th sample, and y_i is the target value of the ith sample.

(c) Calculate the squared error of each sample:

$$e_{ti} = \frac{\left(y_i - h_t\left(x_i\right)\right)^2}{E_t^{\,2}}. \tag{14}$$

(d) Calculate the regression error rate of the base learner h_t in the training data set:

$$\varepsilon_t = \sum_{i=1}^{m} \omega_{ti} e_{ti}. \tag{15}$$

(e) Calculate the weight coefficient of the base learner h_t:

$$\alpha_t = \frac{\varepsilon_t}{1 - \varepsilon_t}. \tag{16}$$

(f) Update the sample distribution of the training set:

$$D_{t+1} = \left(\omega_{t+1,1}, \omega_{t+1,2}, \cdots, \omega_{t+1,m}\right), \tag{17}$$

$$\omega_{t+1,i} = \frac{\omega_{t,i}}{Z_t} \alpha_t^{1-e_{ti}}, i = 1, 2, \cdots, m. \tag{18}$$

In the formula, Z_t represents the normalization factor $Z_t = \sum_{i=1}^{m} \omega_{ti} \alpha_i^{1-e_{ti}}$.

After multiple iterations, a linear combination of learners can be constructed to obtain the final strong learner:

$$H(x) = \sum_{t=1}^{T} ln\left(\frac{1}{\alpha_t}\right) h_t(x). \tag{19}$$

The calculation steps of this combined model are shown in Table 1.

The flow chart is shown in Fig. 3.

In addition to the XGBoost–WSVR–AdaBoost, the RF–WSVR–AdaBoost was also used in this study. The difference between these two combined models is only that the machine learning methods used to obtain the importance scores of each feature are different, but the prediction framework is similar, both using feature weighting ideas.

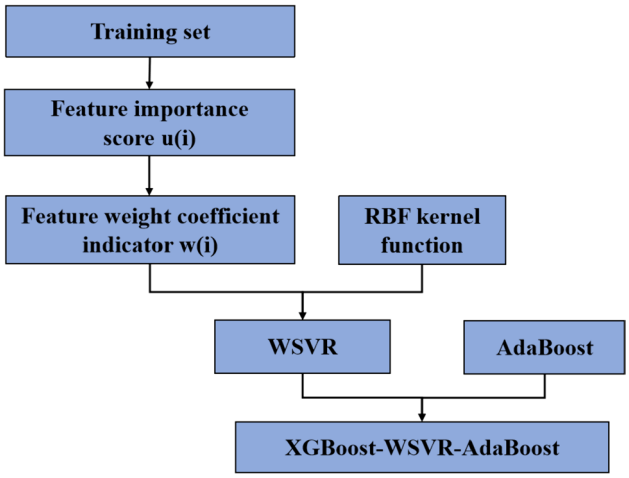

Fig. 3 Algorithm flow chart of XGBoost–WSVR–AdaBoost

Table 1 Algorithm of XGBoost–WSVR–AdaBoost

Algorithm of XGBoost–WSVR–AdaBoost
Input: training set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, where,$x_i \in R^n$,$y_i \in R$; training round T
1. Build the XGBoost model and obtain the importance score u_i of each feature;
2. Sort the feature importance scores u_i to obtain the ranking v_i of each feature variable;
3. Construct the feature weight coefficient index ω_i;
4. Construct the diagonal matrix P of ω_i;
5. Construct feature weighted RBF kernel function: $K_{p-RBF}(x_i, x_j)$;
6. Construct an SVR using the feature weighted RBF kernel function, that is, WSVR, denoted as ζ;
7. Construct the AdaBoost model based on ζ and obtain the final strong learner $H(x)$;
Output: $H(x)$

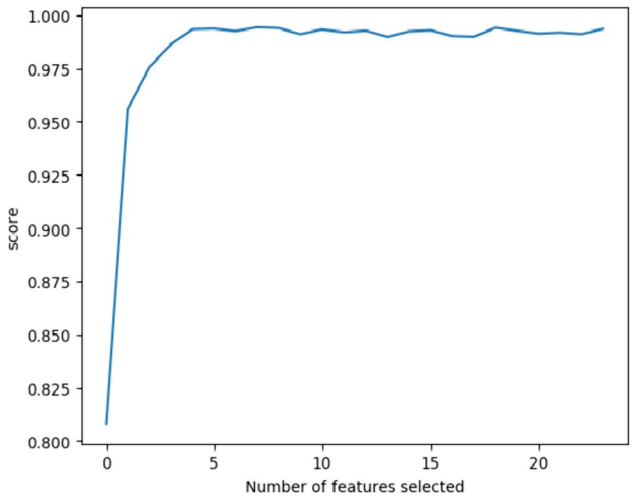

Fig. 4 Average REFCV score for feature subsets with different number of features

3 Numerical Experiments

3.1 Data Sources and Feature Engineering

3.1.1 Financial Data

The financial data used in this study come from Eastmoney.com. The time range of the data is from January 2, 2020 to March 20, 2023 (excluding market non-operating hours), and the data frequency is daily. There are 63 characteristic data in total, including semiconductor industry-related indicators, semiconductor upstream raw material-related indicators, and semiconductor downstream enterprise-related indicators. After obtaining the data, random forests are first used to fill in missing values, and then all filled data are normalized to avoid the impact of different data units or dimensions. Then box plots are used to perform outlier testing and it is found that the data has no Outliers. Next, feature screening is carried out in two steps: first, using the filtering method to delete feature variables with a Pearson correlation coefficient less than 0.5 with the PHLX Semiconductor Sector (SOX), leaving 23 financial features after deletion; second, using the wrapping method, applying RFECV [REF (Recursive Feature Elimination) and CV (Cross Validation)] performs further feature screening. When the number of features is 7, the corresponding feature subset has the highest score, as shown in Fig. 4.

Therefore, this article believes that the optimal subset contains 7 features. Through further analysis, we can see that these seven characteristics are the Taiwan semiconductor return index, SONY closing price (CNY), ST closing price (CNY), AMD closing price (CNY), TXN closing price (CNY), QCOM closing price (CNY), and DMC index of organic silicon. The partial data of various financial indicators and the PHLX Semiconductor Sector (SOX) are shown in Table 2.

Table 2 Results of some financial indicators

Indicator name	SOX	Taiwan Semiconductor Return Index	SONY Closing Price (CNY)	…	QCOM Closing Price (CNY)	DMC Index of Organic Silicon
2020-01-02	0.22	0.20	0.23	…	0.22	0.08
2020-01-03	0.21	0.20	0.21	…	0.20	0.08
2020-01-06	0.20	0.19	0.23	…	0.20	0.08
…	…	…	…	…	…	…
2023-03-16	0.66	0.67	0.45	…	0.46	0.04
2023-03-17	0.65	0.70	0.45	…	0.47	0.04
2023-03-20	0.66	0.69	0.48	…	0.47	0.04

Table 3 Cleaned text data

Date	Newsheadlines
2020/1/4	Guangdong, speed up, semiconductor, IC, industry, development, cultivation, modern, industry, cluster
2020/1/4	Hubei, university, the first, chip, industry, college, unveiling, establishment
2020/1/6	China, Fund, Zongting Zhao, rise, chip, the best of the best
…	…
2023/3/20	Scientific and technological innovation, index, rise, chip, plate, performance
2023/3/20	Semiconductor, plate, once again, active, many stocks, rise
2023/3/20	Boe, investment, establishment, digital, technology, new company, semiconductor, lighting, devices, sales business

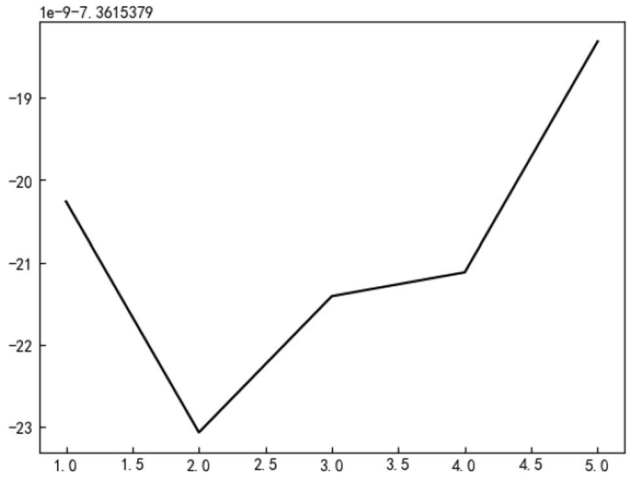

Fig. 5 Perplexity in the number of different topics

3.1.2 Text Data

This paper used Python to crawl a total of 29,211 news headlines about semiconductors, chips, and integrated circuits from January 4, 2020, to March 20, 2023, on Eastmoney.com. After acquiring news headlines, text denoising, word segmentation, and stop-word removal were carried out. The results after text data cleaning are shown in Table 3.

Then, the LDA topic model was applied to further retrieve the information contained in different topics. In this paper, the confusion degree was applied to determine the number of LDA topics, and it was found that the confusion degree was the lowest when the number of topics was two. As shown in Fig. 5, the horizontal axis represents the number of topics, and the vertical axis represents the perplexity.

After the number of themes is determined, the Gensim library of Python imports the preprocessed network news headlines into the LDA model, and the probability of each news under each topic can be calculated. By dividing each news into the theme with large probability value, the theme distribution of each news can be obtained. Results for the distribution of news themes are shown in Table 4.

Then the SnowNLP library in Python is used to calculate the emotion score of each news item. News with an emotion score greater than 0.5 is marked as 1, indicating positive emotion; a value less than 0.5 is marked as 0, indicating negative emotion; and a value equal to 0.5 is marked as 2, with no obvious emotional tendencies. The emotion score and labels of news are shown in Table 5.

By dividing each news headline according to the time interval of 1 day, the number of daily news M_t, the number

Table 4 Topic distribution of news headlines

Date	News headlines	Topics1	Topics2	Topic
2020/1/4	Guangdong, speed up, semiconductor, IC, industry, development, cultivation, modern, industry, cluster	0.04366779	0.956332445	2
2020/1/4	Hubei, university, the first, chip, industry, college, unveiling, establishment	0.06167011	0.938337922	2
2020/1/6	China, Fund, Zongting Zhao, rise, chip, the best of the best	0.476152778	0.5246647	2
…	…	…	…	…
2023/3/20	Scientific and technological innovation, index, rise, chip, plate, performance	0.935506463	0.064495012	1
2023/3/20	Semiconductor, plate, once again, active, many stocks, rise	0.926127553	0.073871575	1
2023/3/20	Boe, investment, establishment, digital, technology, new company, semiconductor, lighting, devices, sales business	0.040303696	0.959703445	2

Table 5 Sentiment score and labels

Date	News headlines	Sentiment	Tag
2020/1/4	Guangdong, speed up, semiconductor, IC, industry, development, cultivation, modern, industry, cluster	0.997945079	1
2020/1/4	Hubei, university, the first, chip, industry, college, unveiling, establishment	0.820975059	1
2020/1/6	China, Fund, Zongting Zhao, rise, chip, the best of the best	0.999457968	1
…	…	…	…
2023/3/20	Scientific and technological innovation, index, rise, chip, plate, performance	0.995878178	1
2023/3/20	Semiconductor, plate, once again, active, many stocks, rise	0.821809752	1
2023/3/20	Boe, investment, establishment, digital, technology, new company, semiconductor, lighting, devices, sales business	0.941576633	1

Table 6 Text characteristics

Date	M_t	M_t^{pos}	M_t^{neg}	mean_senti	Topic1	Topic2	B_t	B_t^*	B_t^{Att}
2020/1/4	3	3	0	0.94	0.00	0.94	1	1.39	1.39
2020/1/6	2	2	0	1.00	0.00	1.00	1	1.10	1.10
2020/1/9	1	0	1	0.30	0.00	0.30	−1	−0.69	−0.69
…	…	…	…	…	…	…	…	…	…
2023/3/21	10	9	1	0.85	0.80	1.61	1.92	0.83	0.92
2023/3/22	13	12	1	0.89	0.85	1.87	2.23	0.80	0.90
2023/3/23	71	54	17	0.76	0.52	1.12	2.23	0.80	0.74

of daily positive emotion news M_t^{pos}, and the number of daily negative emotion news M_t^{neg} can be obtained. The average daily emotion score (mean_senti) and the average daily emotion score under various topics (topic1, topic2) can also be calculated. In addition, B_t, B_t^*, B_t^{Att} were also introduced.

Finally, the text features of this article were constructed, and some of the results are shown in Table 6.

To match the time series data in the empirical analysis and ensure their consistency. first, the text data on non-trading days were deleted; second, random forest was used to fill in missing values, and all data were normalized; then, a boxplot was used to test and deal with outliers; finally, the above text features were subjected to a 1–5 order lag, and then RFECV was applied to screen the features to generate the highest-precision text feature subset. This subset contains two text features, namely, the average emotion score under topic 2 with a fifth-order lag and B_t^{Att} with a fifth-order lag.

The trend of the PHLX Semiconductor Sector (SOX), the average emotion score under topic 2 with a fifth-order lag, and B_t^{Att} with a fifth-order lag are shown in Fig. 6.

From the change trends in Fig. 6, it can be seen that in the early stage, with the steady increase in the SOX, bt_att_Lag5 rises from a lower level to a higher level, and topic2_ Lag5 is mostly above 0.5 points. When the SOX fluctuates, bt_att_Lag5 and topic2_ Lag5 also show a sharp and significant fluctuation trend. From this, it can be seen that bt_att_Lag5 and topic2_ Lag5 can reflect economic behavior, so this article believes that combining them with economic characteristics can help predict the SOX.

3.1.3 Google Trends Data

This paper used Google Trends as a proxy indicator to measure investor attention. Keywords related to semiconductors and the daily frequency data of Google Trends for those keywords were used as search engine data. Four specific Google Trends were selected: "semiconductor", "semiconductor stock", "chip price", and "Semiconductor Price". Data from Google Trends (http://www.google.com/trends) were obtained in the period from January 1, 2020, to March 31, 2023.

This paper referred to Xu et al.'s proposed method (2018) [37] to process the Google Trends data. Compare adjusted daily frequency Google Trends, weekly frequency Google Trends which calculated based on the adjusted daily frequency Google Trends and the actual weekly frequency Google Trends which provided by Google. The comparison of three time series for the four Google Trends (semiconductor, semiconductor stock, chip price, and semiconductor price) is shown in Fig. 7.

The calculated weekly Google Trends was very close to the actual weekly Google Trends, which proves that our method was effective for calculating daily Google Trends throughout the entire time range.

After obtaining the adjusted Google Trends data, the non-trading day Google Trends data were first deleted to ensure the consistency of the time series data. Secondly, random forest was used to fill in missing values and all data were normalized. Then, a boxplot was used to test and deal with outliers. The partial data after pretreatment are shown in Table 7.

Finally, the four Google Trends features were subjected to a 1–5-order lag, and then RFECV was applied to screen the features. The highest-precision subset of Google Trends features contains three Google Trends features: semiconductor (global) with a two-step lag, semiconductor stock (global) with a five-step lag, and chip price (global) with a one-step lag.

The time series data of the SOX, semiconductor(global) with a two-step lag, semiconductor stock(global) with a five-step lag and chip price(global) with a one-step lag are shown in Fig. 8.

Figure 8 showsreal value of the testreal value of the test interesting findings. The trend of the PHLX Semiconductor Sector (SOX) fluctuates significantly throughout the entire range. Relatively speaking, the Google Trends is relatively stable in trend, but the frequency of fluctuations is relatively high. Meanwhile, the fluctuation behavior of Google Trends with different keywords has certain similarities. There is a certain correlation between investor attention and the PHLX Semiconductor Sector (SOX). When the PHLX Semiconductor Sector (SOX) rises, attention also increases. When the PHLX Semiconductor

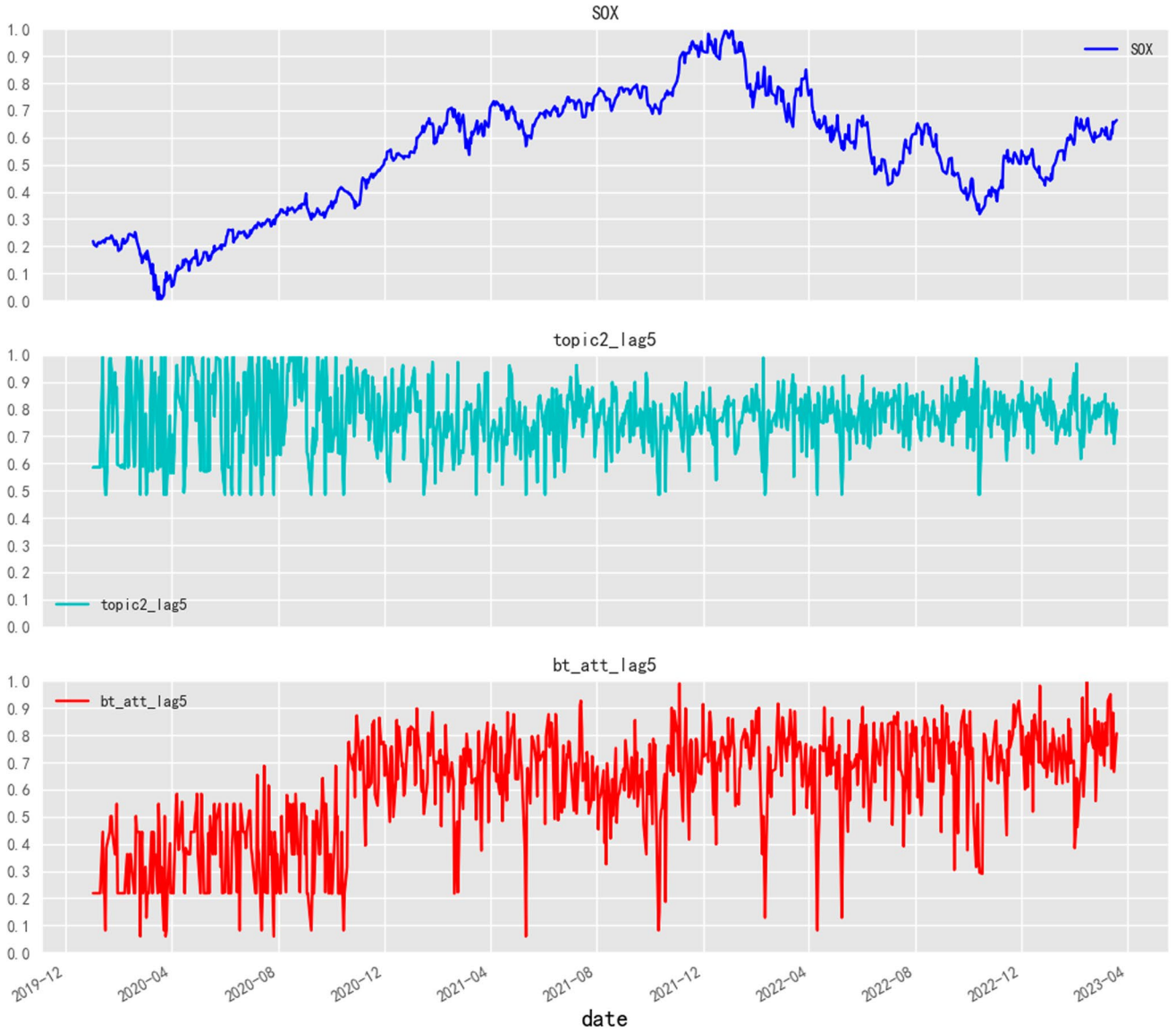

Fig. 6 Comparison chart of change trends

Sector (SOX) decreases, attention decreases. When the PHLX Semiconductor Sector (SOX) fluctuates, attention also fluctuates significantly. This means that the information contained in Google Trends may help predict the PHLX Semiconductor Sector (SOX). The possible reason why Google Trends may be a promising predictor of the PHLX Semiconductor Sector (SOX) is that Google Trends directly reflect public attention to the semiconductor industry, and due to the herd effect, it is likely to affect semiconductor trends. Next, this article will further predict and study the PHLX Semiconductor Sector (SOX) using Google Trends indicators.

3.2 Evaluation Index and Parameter Selection Methods of Each Model

The mean absolute percentage error (MAPE) and coefficient of determination (R^2) were used to measure the performance of different forecasting methods, where MAPE represents the mean absolute percentage error, the influence of the data range size is avoided by calculating the percentage error between the true value and the prediction, and the smaller the value of the MAPE, the better the accuracy of the prediction model. R^2 is the coefficient of determination, and the larger the R^2, the better the model fitting effect. The calculation formula is as follows, where n denotes the number of test set samples,y_i denotes the predicted value of the test set, and x_i denotes the real value of the test set:

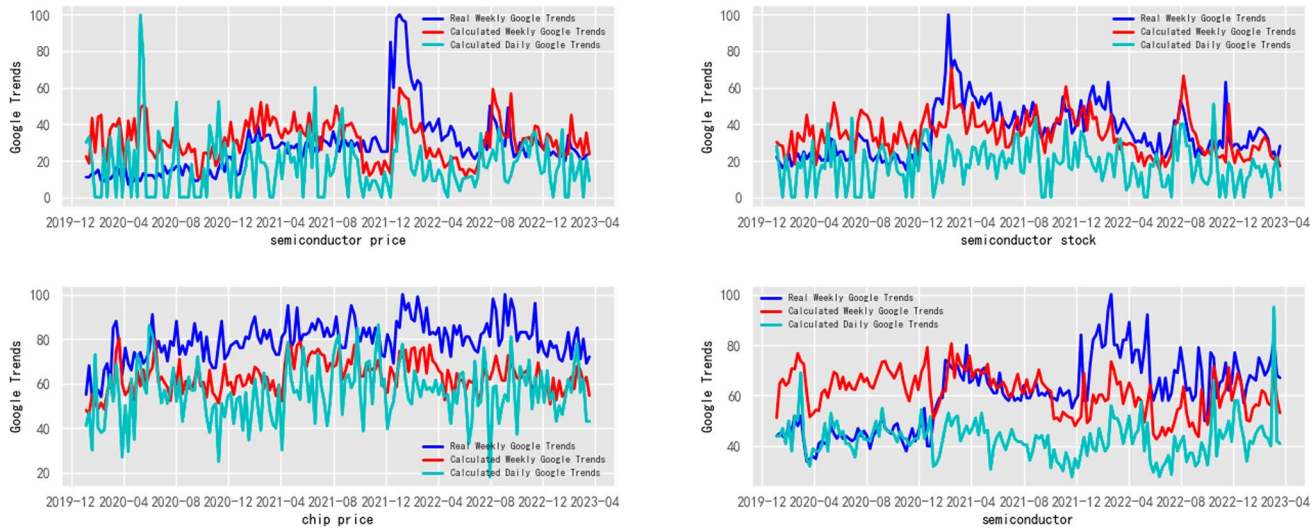

Fig. 7 Comparison chart of different time series

Table 7 Pretreated google trends

Indicator name	SOX	Semiconductor	Semiconductor stock	Chip price	Semiconductor price
2020/1/2	0.2184	0.2208	0.6785	0.4519	0.4988
2020/1/3	0.206	0.4896	0.3209	0.5151	0.2702
2020/1/6	0.199	0.54	0.1268	0.4392	0.4586
…	…	…	…	…	…
2023/3/16	0.658	0.288	0.2188	0.4266	0.2494
2023/3/17	0.6527	0.1704	0.1983	0.376	0.1455
2023/3/20	0.664	0.3552	0.2801	0.4772	0.4884

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right|}{n}, \tag{20}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{21}$$

Machine learning algorithms require user-defined inputs to achieve a balance between accuracy and generality, a process called hyperparameter tuning, also known as hyperparameter optimization. Models need to search for the best configuration of hyperparameters to achieve the best performance. There are many tools and methods for adjusting hyperparameters, such as coordinate descent, GridSearchCV, RandomizedSearchCV, and Bayesian optimization. Because GridSearchCV can be automatically adjusted within the specified range, the optimized results and parameters can be obtained by simply inputting the parameters, which is time-saving, labor-saving, concise, and flexible. Therefore, the mesh search method was used in the modeling process of each model.

3.3 Empirical Results and Evaluation

This paper used data from January 2, 2020, to July 27, 2022, as the training set, and data from July 27, 2022, to March 20, 2023, as the test set, as shown in Fig. 9.

12 models including Adaptive Boosting (AdaBoost), lasso regression, a recurrent neural network (RNN), extremely randomized forest, (multi-layer perceptron) MLP, support vector regression, XGBoost–WSVR, long–short-term memory (LSTM), SVR–AdaBoost, ridge regression, RF–WSVR–AdaBoost and XGBoost–WSVR–AdaBoost were used to predict the PHLX Semiconductor Sector (SOX).

Due to the normalization of various variables in the process of feature engineering, to make the evaluation results more reasonable, after the prediction results were obtained,

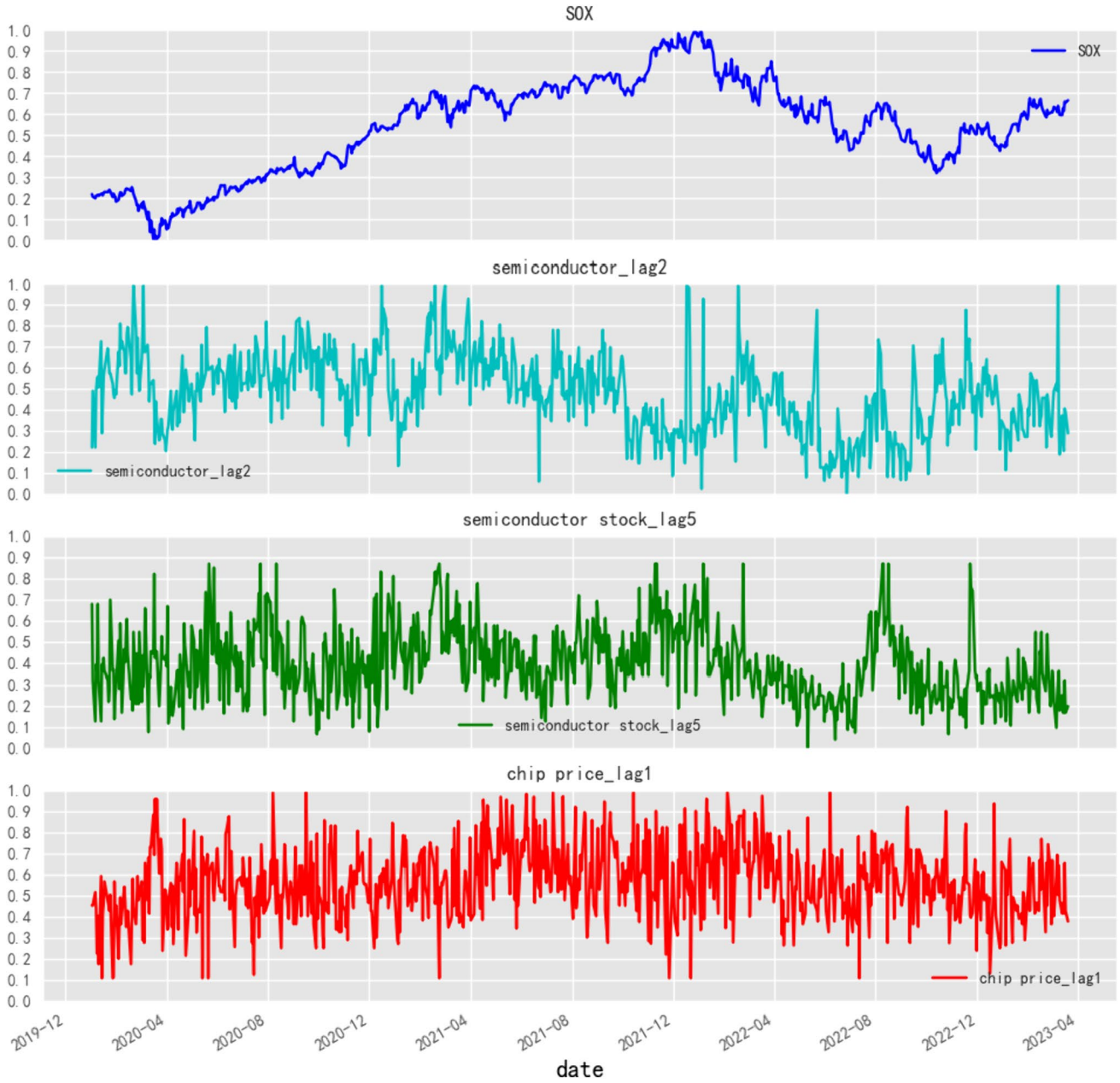

Fig. 8 Google Trends condicators trend charts

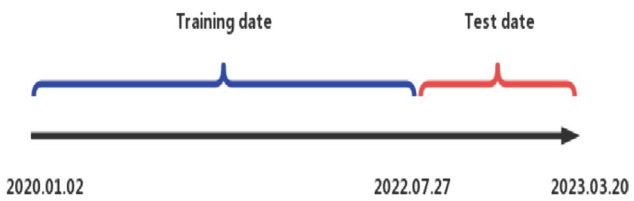

Fig. 9 Partition of training and testing sets

the real value and predicted value of the target variable were first processed via reverse normalization, and then the fitting effect of each model was evaluated via comparative analysis.

The predicted results of the PHLX Semiconductor Sector (SOX) for each model using three sources of data are shown in the following table. Where y_test is the true value on the test set, and AdaBoost, Lasso, etc., are the predicted values on the test set under each model. Table 8 shows the prediction results using only the financial indicators, Table 9 shows the prediction results combining the text indicators,

Table 8 Prediction results using only the financial indicators

Date	y_test	AdaBoost	Lasso	…	Ridge	RF–WSVR–AdaBoost	XGBoost–WSVR–AdaBoost
2022/7/28	2944.50	2911.95	2940.96	…	2899.35	2974.16	2973.78
2022/7/29	2967.07	2918.99	2969.63	…	2924.56	3003.92	3005.03
2022/8/1	2978.36	2918.99	2977.87	…	2941.31	3027.63	3028.64
2022/8/2	2974.78	2918.51	2945.42	…	2922.61	3010.65	3014.63
2022/8/3	3053.50	2938.72	3036.39	…	2991.45	3085.16	3085.09
…	…	…	…	…	…	…	…
2023/3/14	3010.29	2900.38	2958.98	…	2934.40	3009.39	3005.09
2023/3/15	2977.25	2900.38	2922.35	…	2899.91	2976.79	2973.70
2023/3/16	3098.10	2906.01	2982.82	…	2982.55	3073.86	3068.84
2023/3/17	3083.51	2900.38	2978.91	…	2986.34	3078.44	3071.88
2023/3/20	3114.61	2906.01	3026.89	…	3022.78	3115.64	3106.77

Table 9 Prediction results combining the text indicators

Date	y_test	AdaBoost	Lasso	…	Ridge	RF–WSVR–AdaBoost	XGBoost–WSVR–AdaBoost
2022/7/28	2944.50	2879.26	2938.54	…	2906.78	2959.74	2953.11
2022/7/29	2967.07	2879.26	2963.72	…	2940.65	2994.46	2990.00
2022/8/1	2978.36	2879.26	2973.77	…	2944.89	3008.10	3001.81
2022/8/2	2974.78	2879.26	2942.82	…	2940.28	3002.36	2997.10
2022/8/3	3053.50	2879.26	3031.55	…	2991.52	3057.32	3051.84
…	…	…	…	…	…	…	…
2023/3/14	3010.29	2797.02	2948.67	…	2940.73	2968.24	2978.25
2023/3/15	2977.25	2817.39	2911.73	…	2910.56	2940.93	2950.67
2023/3/16	3098.10	2887.98	2975.95	…	3005.87	3039.84	3049.15
2023/3/17	3083.51	2890.30	2972.67	…	2993.46	3028.24	3035.82
2023/3/20	3114.61	2892.09	3020.57	…	3037.18	3070.87	3079.23

Table 10 Prediction results combining the Google trends indicators

Date	y_test	AdaBoost	Lasso	…	Ridge	RF–WSVR–AdaBoost	XGBoost–WSVR–AdaBoost
2022/7/28	2944.50	2927.46	2939.09	…	2955.11	2957.76	2966.63
2022/7/29	2967.07	2954.68	2965.40	…	2945.81	2962.38	2977.44
2022/8/1	2978.36	3019.08	2976.85	…	2984.15	2997.91	3013.36
2022/8/2	2974.78	2918.51	2948.42	…	2962.09	2997.24	3008.61
2022/8/3	3053.50	3083.78	3034.39	…	3035.56	3053.13	3066.98
…	…	…	…	…	…	…	…
2023/3/14	3010.29	2910.92	2936.24	…	2979.05	2953.66	2972.17
2023/3/15	2977.25	2910.92	2903.39	…	2943.17	2922.14	2938.80
2023/3/16	3098.10	3044.56	2971.66	…	3032.90	3026.14	3048.95
2023/3/17	3083.51	3044.56	2969.78	…	3014.86	3009.02	3029.91
2023/3/20	3114.61	3044.56	3015.96	…	3059.93	3044.13	3063.90

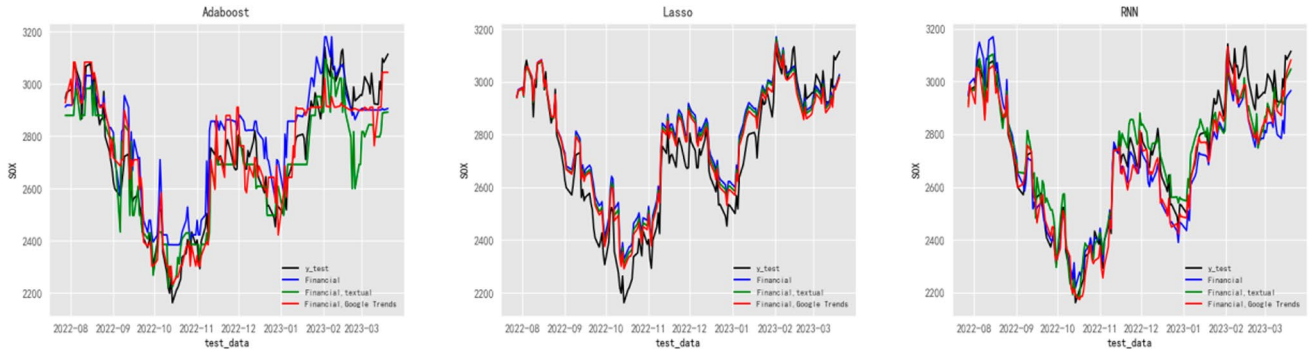

Fig. 10 Prediction results of AdaBoost, Lasso, RNN

and Table 10 shows the prediction results combining the Google trends indicators.

The predicted results of the PHLX Semiconductor Sector (SOX) for each model using three sources of data are shown in the following figures. Among them, the black line represents the true value of the PHLX Semiconductor Sector (SOX) in the test set, the blue line represents the predicted results based on financial indicators, the green line represents the predicted results combined with text indicators, and the red line represents the predicted results combined with Google Trends indicators.

As shown in Fig. 10, for Adaboost, whether combining text indicators or Google Trends indicators, it can solve the problem of large deviation in extreme value prediction when only using financial indicators in the early stage. However, in February and March 2023, combining text indicators or Google features indicators worse and widened the distance from the true value.

In Lasso regression, combining text indicators or Google Trends indicators can better predict the PHLX Semiconductor Sector (SOX), and the predicted results are closer to the real value. Compared to combining text indicators, the effect of getting closer to the real value is more significant when combining Google Trends indicators. Which also indicates that Google Trends has a better improvement in predicting the PHLX Semiconductor Sector (SOX).

About RNN, combining text indicators or Google Trends indicators can better predict the true value most of the time, and adding Google Trends indicators has a more significant effect. However, in November 2022, adding text indicators or Google Trends indicators to predictions showed worse results, deviating more from the true value compared to using only financial indicators.

As shown in Fig. 11, for extreme random forests, it can be observed from the graph that the prediction results with the addition of text indicators or Google Trends indicators in the early stage are not significantly different from those with only financial indicators. However, in February and March 2023, the prediction effect showed a significant improvement.

For MLP, the addition of text indicators or Google trends indicators can significantly improve the prediction effect, but after February 2023, it increases the distance with the real value.

For SVR, within the entire prediction range, the difference in prediction performance among the three sources of data is relatively small. Adding text indicators or Google Trends indicators only shows a small improvement in prediction performance.

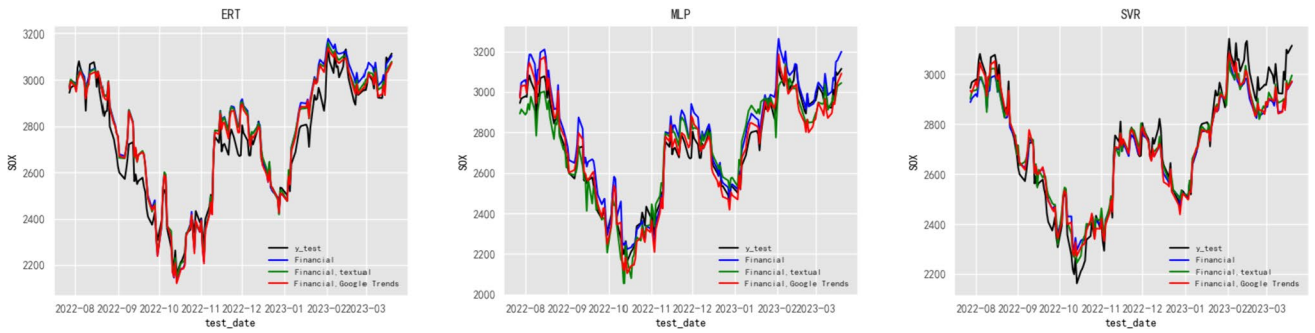

Fig. 11 Prediction results of ERT, MLP, SVR

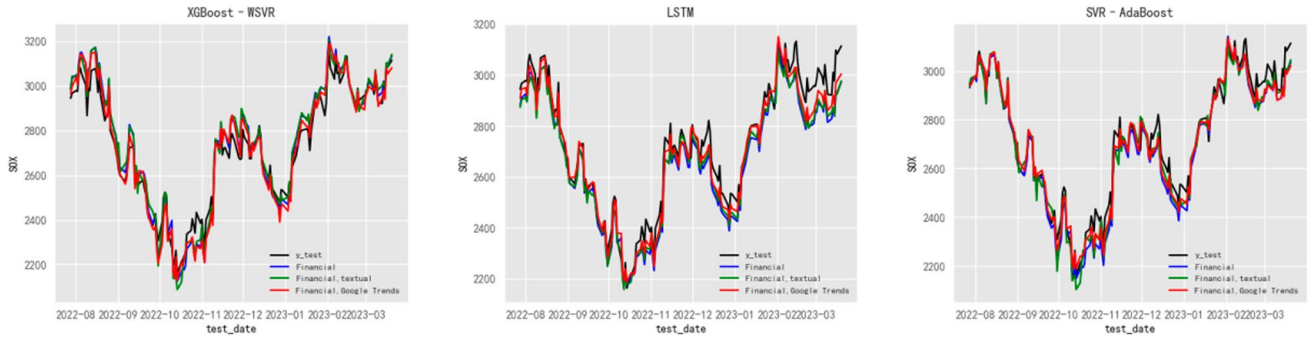

Fig. 12 Prediction results of XGBoost–WSVR, LSTM, SVR–AdaBoost

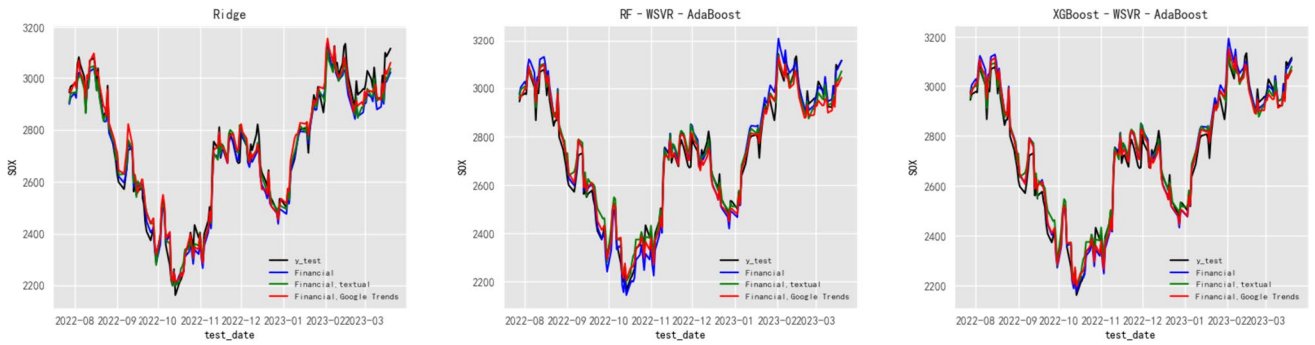

Fig. 13 Prediction results of Ridge, RF–SVR–AdaBoost, XGBoost–SVR–AdaBoost

As shown in Fig. 12, for XGBoost–WSVR, the predicted value is significantly smaller at the minimum and significantly larger at the maximum. Relatively speaking, adding Google trends indicator is better than adding text indicator, and the predicted value is closer to the true value.

For LSTM, there is almost no difference in the prediction effect of the three sources in the early stage, but since November 2022, the prediction effect of adding text index or Google trends index has been significantly improved, and the forecast value is closer to the true value.

For SVR-AdaBoost, at the extreme value, the prediction effect of all three data sources needs to be improved. But at minima, adding Google trends indicators significantly narrowed the distance between the predicted value and the true value.

As shown in Fig. 13, it can be seen that for the three models: Ridge, RF–SVR–AdaBoost and XGBoost–SVR–AdaBoost, which already performed well when using only financial indicators. The addition of text indicators or Google trends indicators has improved the prediction effect of PHLX Semiconductor Sector (SOX) to a certain extent, and the forecast value is closer to the true value. For Ridge, the most significant periods were in February 2023 and March 2023, and for RF–SVR–AdaBoost and XGBoost–WSVR–AdaBoost, the most obvious periods were in November 2022 and March 2023.

The MAPE and R^2 of each source data on the test set under each model are shown in Table 11. In addition, Figs. 14 and 15 further show the comparison results of MAPE and R^2, where blue indicates the results using only financial indicators, green indicates the results combined with text indicators, and orange indicates the results of Google trends indicators.

According to the results of Figs. 14, 15 and Table 11, First, Among the 12 models and data from three sources, the XGBoost–WSVR–AdaBoost combined with Google Trends features had the best prediction effect for the PHLX Semiconductor Sector (SOX). The *MAPE* was the smallest (0.0116), and the R^2 was the largest (0.9783). The models with minimum MAPE and maximum R^2 for prediction with only financial indicators and combined with Google trends indicators are all XGBoost–WSVR–AdaBoost. The model with minimum MAPE and maximum R^2 when prediction is combined with text indicators is ridge regression.

Second, comparing 12 models. As it can be found, in addition to the ridge regression combining textual metrics, The XGBoost–WSVR–AdaBoost always performs the best. In contrast to AdaBoost and SVR-AdaBoost, The prediction effect has been significantly improved, the importance of

Table 11 Prediction results of each model based on various source data

Model	MAPE			R^2		
	Financial features	Financial features; textual features	Financial features; google trends features	Financial features	Financial features; textual features	Financial features; google trends features
AdaBoost	0.0380	0.0291	0.0257	0.7934	0.8360	0.8893
Lasso	0.0315	0.0276	0.0249	0.8611	0.8937	0.9136
RNN	0.0251	0.0217	0.0190	0.8895	0.9207	0.9304
ERT	0.0231	0.0202	0.0194	0.9128	0.9283	0.9355
MLP	0.0237	0.0235	0.0182	0.9088	0.9122	0.9479
SVR	0.0200	0.0180	0.0168	0.9312	0.9465	0.9500
XGBoost–WSVR	0.0188	0.0184	0.0159	0.9454	0.9456	0.9567
LSTM	0.0226	0.0204	0.0142	0.9186	0.9312	0.9625
SVR–AdaBoost	0.0168	0.0146	0.0128	0.9538	0.9620	0.9711
Ridge	0.0131	**0.0120**	0.0122	0.9684	**0.9747**	0.9735
RF–WSVR–AdaBoost	0.0144	0.0125	0.0117	0.9681	0.9733	0.9768
XGBoost–WSVR–AdaBoost	**0.0127**	0.0122	**0.0116**	**0.9736**	0.9741	**0.9783**

Bold values indicates to show the best prediction

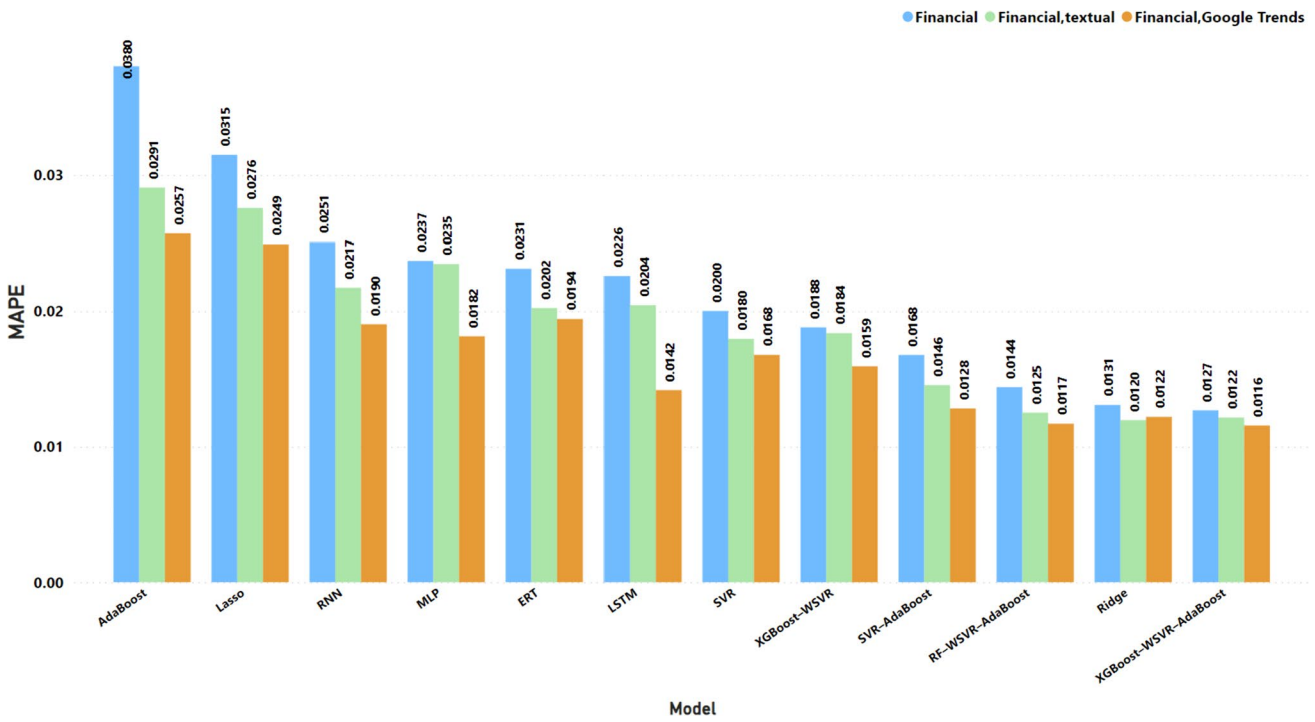

Fig. 14 Comparison results of MAPE

performing the feature weighting is demonstrated. Compared with XGBoost–WSVR, the prediction effect is also significantly improved, proving the importance of iterative updating of AdaBoost to a strong learner. In addition, the prediction effect of RF–WSVR–AdaBoost is also good, second only to XGBoost–WSVR–AdaBoost.

These two models get different feature importance scores, but both adopting the idea of feature weighting, which further proves the importance of the feature weighting prediction framework proposed in this paper and good prediction results.

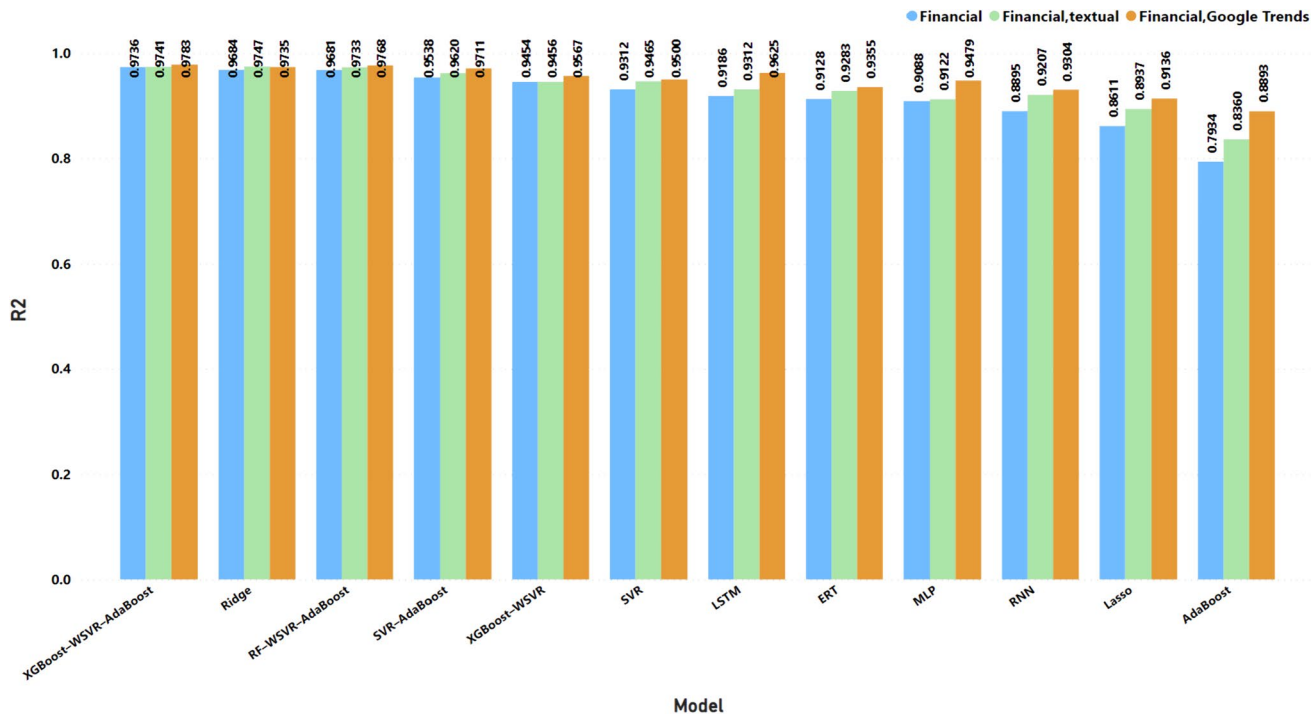

Fig. 15 Comparison results of R^2

Finally, Comparing the data of the three sources. It can be found that the prediction effect of combining financial indicators or combining Google trends indicators is significantly better than that of using financial indicators only. Through further analysis, it can be found that in addition to the prediction effect of ridge regression combined with text index is slightly better than that of Google trends index, combining Google trends index is more obvious for predicting the improvement effect of PHLX Semiconductor Sector (SOX).

4 Conclusions

4.1 Research Conclusion

In this paper, we construct a feature weight coefficient index is constructed, and on this basis, the combined model XGBoost–WSVR–AdaBoost is constructed. In terms of data selection, this paper not only uses traditional financial data, but also introduces search engine data (Google Trends) and text data (network news headlines). On this basis, the PHLX Semiconductor Sector (SOX) prediction model is constructed based on financial characteristics, text features and Google trends features. 11 benchmark models were introduced for comparative analysis. The following conclusions can be obtained:

First, among the 12 models and data from three sources, the XGBoost–WSVR–AdaBoost combined with Google Trends features had the best prediction effect for the PHLX Semiconductor Sector (SOX). The *MAPE* was the smallest (0.0116), and the R^2 was the largest (0.9783).

Second, comparing the 12 models. In addition to the ridge regression when combining textual indicators, the proposed combined model XGBoost–WSVR–AdaBoost based on feature-weighted features always performs the best. Compared with the traditional AdaBoost, XGBoost–WSVR–AdaBoost significantly improved the PHLX Semiconductor Sector (SOX), with R^2 from the three sources increasing from 0.7934, 0.8360 and 0.8893 to 0.9736, 0.9741 and 0.9783, respectively. Moreover, compared with SVR–AdaBoost, XGBoost–WSVR and RF–WSVR–AdaBoost, XGBoost–WSVR–AdaBoost predicted better, which proves the effectiveness of feature-weighting AdaBoost. At the same time, the prediction effect of RF–WSVR–AdaBoost is also good, second only to XGBoost–WSVR–AdaBoost. These two models are get different feature importance scores, but both adopting the idea of feature weighting, which further proves the importance of the feature weighting prediction framework proposed in this paper and good prediction results.

Finally, comparing the data of the three sources. It can be found that the prediction effect of combining financial indicators or combining Google trends indicators is significantly better than that of using financial indicators

only. On 12 models, based on financial characteristics, combined with text characteristics, combined with Google trends of PHLX Semiconductor Sector (SOX) forecast the average of R^2 is 0.9187, 0.9332, 0.9488, respectively, the average of MAPE is 0.0217, 0.0192, 0.0169. In addition to the prediction effect of ridge regression when combining with text index is slightly better than that of Google trends index, compared with other models, the improvement effect of Google trends index is more obvious in predicting the improvement effect of PHLX Semiconductor Sector (SOX). From the analysis of data dimensions of different sources, it can be found that adding news text emotion and investor attention is very effective for predicting the PHLX Semiconductor Sector (SOX), which generally improves the effect of the prediction model. And compared with the news text emotion, introducing investor attention is more effective.

4.2 Significance of the Study

It is important to predict semiconductor prices more accurately:

First of all, at the national level. The semiconductor industry belongs to the high-tech industry, which is the key development goal in the future, and is related to the development of the future scientific and technological strength. Semiconductor price forecast can provide a reference for the policy formulation of government departments and promote the development of semiconductor industry.

Secondly, at the enterprise level. Semiconductor price prediction helps semiconductor enterprises to better understand the market operation rules, understand the market dynamics of the production and sales of semiconductor enterprises; helps enterprises to formulate more scientific and reasonable strategic planning, improve the operating efficiency of enterprises, and further promote the stability and development of the market.

Finally, at the level of market participants. Through the analysis and prediction of the operating conditions, industry trends, market demand and competitive factors of semiconductor companies. It can help market participants better understand the future trends of the industry and the company, and improve their decision-making ability. Help investors to better plan their investment strategies to obtain higher investment returns.

4.3 Shortcomings and Prospects

The prediction of price index in semiconductor industry is an interesting and meaningful research direction. This paper mainly studies the prediction of time series data from the perspective of combining network big data, text data and combined model, and provides some ideas and conclusions. However, there are still some limitations and deficiencies in the study. In future studies, some improvements can be considered from the following points to improve the prediction performance.

(1) More text data sources can be considered, such as semiconductor-related online reviews, semiconductor-related literature summary, and summary of international semiconductor-related conferences. More models can also be considered in the refinement of emotion scores;
(2) In addition to Google Trends, more types of network big data, such as Baidu Trend. And you can select more keywords and combine them to build related Google trends and other network big data indicators;
(3) More neural network models can be considered in the prediction model, and other kernel functions can be tried to improve, which may get better results.

Author Contributions Feng Chen: conceptualization, methodology, reviewing, and editing; Qi Jiang: data curation, writing—original draft preparation and software; Hongyu Deng: validation and writing—reviewing and editing.

Funding This work was supported by the National Natural Science Foundation of China (Grant no. 11701077) and the Natural Science Foundation of Jilin Province (Grant no. 20220101026JC, 20220201160GX, and 20210101476JC).

Data Availability The financial data and text data used in this study were from Eastmoney.com. Google trends data were from Google Trends (http://www.google.com/trends).

Declarations

Conflict of Interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Wu, X., Zhang, X., Shen, H.: Business model innovation in China's semiconductor industry: a pathway to technological breakthrough. Front. Bus. Res. China **17**(4), 467–497 (2023)
2. Tutsoy, O.: Graph theory based large-scale machine learning with multi-dimensional constrained optimization approaches for exact epidemiological modelling of pandemic diseases. IEEE Trans. Pattern Anal. Mach. Intell. **45**(8), 9836–9845 (2023)
3. Chen, H., Lin, M., Liu, J., Yang, H., Zhang, C., Xu, Z.: NT-DPTC: a non-negative temporal dimension preserved tensor completion model for missing traffic data imputation. Inf. Sci. **653**, 119797 (2024)
4. Ong, C.S., Huang, J.J., Tzeng, G.H.: Model identification of ARIMA family using genetic algorithms. Appl. Math. Comput. **164**(3), 885–912 (2005)
5. Chen, T.: A collaborative fuzzy-neural approach for forecasting the price of a DRAM product. Int. J. Technol. Intell. Planning **7**(2), 95–109 (2011)
6. Xu, Q.: Research on stock selection strategy in Global semiconductor Industry based on MLP Neural Network. Master's thesis, Shanghai Normal University. https://doi.org/10.27312/d.cnki.gshsu.2020.000677(2020) (**in Chinese**)
7. Grau, I., de Hoop, M., Glaser, A., Nápoles, G., & Dijkman, R.: Semiconductor Demand Forecasting using Long Short-term Cognitive Networks. In 34th Benelux Conference on Artificial Intelligence and 31st Belgian-Dutch Conference on Machine Learning, BNAIC/BeNeLearn 2022. Antwerpen University (2022).
8. Munirathinam, S., & Ramadoss, B.: Big data predictive analtyics for proactive semiconductor equipment maintenance. In 2014 IEEE International Conference on Big Data (Big Data) .pp. 893–902. IEEE (2014)
9. Ochonogor, K.N., Osho, G.S., Anoka, C.O., Ojumu, O.: The COVID-19 pandemic and supply chain disruption: an analysis of the semiconductor industry's resilience. Int. J. Tech. Sci. Res. Eng. **6**(1), 7–18 (2023)
10. Chien, C.F., Ehm, H., Fowler, J.W., Kempf, K.G., Mönch, L., Wu, C.H.: Production-level artificial intelligence applications in semiconductor supply chains. IEEE Trans. Semicond. Manuf. **36**(4), 560–569 (2023)
11. Singgih, G.M., Nugraha, E.S.: Forecasting the monthly stock price per share of Taiwan Semiconductor Manufacturing Company Limited (TSM) using ARIMA Box-Jenkins Method. J. Actuarial Finance Risk Manage. **2**(1), 1–9 (2023)
12. Yao, Z.: Research on the impact of investor sentiment on semiconductor sector yields. Master's thesis, Hebei University of Geosciences. https://doi.org/10.27752/d.cnki.gsjzj.2022.000156(2022). (**in Chinese**)
13. Huang, Y.: The long-term relationship between USD to CNY exchange rate and China's semiconductor: an empirical research. Highlights Bus. Econ. Manage. **5**, 156–164 (2023)
14. Chen, T.C.T., Wu, H.C.: Forecasting the unit cost of a DRAM product using a layered partial-consensus fuzzy collaborative forecasting approach. Complex Intell. Syst. **6**, 479–492 (2020)
15. Chen, T.: A hybrid fuzzy and neural approach for DRAM price forecasting. Comput. Ind. **62**(2), 196–204 (2011)
16. Li, Y., Bu, H., Li, J., Wu, J.: The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. Int. J. Forecast. **36**(4), 1541–1562 (2020)
17. Ye, T.: Stock forecasting method based on wavelet analysis and ARIMA-SVR model. In 2017 3rd international conference on information management (ICIM) ,pp. 102–106. IEEE (2017)
18. Vo, A.H., Nguyen, T., Le, T.: Brent oil price prediction using bi-LSTM network. Intell. Autom. Soft Comput. **26**(4), 1307–1317 (2020)
19. Huang, J.Y., Tung, C.L., Lin, W.Z.: Using social network sentiment analysis and genetic algorithm to improve the stock prediction accuracy of the deep learning-based approach. Int. J. Comput. Intell. Syst. **16**(1), 93 (2023)
20. Xia, M., Shao, H., Ma, X., de Silva, C.W.: A stacked GRU-RNN-based approach for predicting renewable energy and electricity load for smart grid operation. IEEE Trans. Industr. Inf. **17**(10), 7050–7059 (2021)
21. Wang, Y., Guo, Y.: Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Commun. **17**(3), 205–221 (2020)
22. Tutsoy, O., Tanrikulu, C.: A machine learning-based 10 years ahead prediction of departing foreign visitors by reasons: a case on Türkiye. Appl. Sci. **12**(21), 11163 (2022)
23. Cowles, A., 3rd.: Can stock market forecasters forecast? Econometrica J. Econ. Soc. **1**(3), 309–324 (1933)
24. Li, C., Zhu, S., Wei, M., Yu, J., Li, X.: Analysis of the emotional tendency of news text based on dictionary and rules. Shandong Sci. **30**(01), 115–121 (2017). (**in Chinese**)
25. Wu, J., Shao, D., Jiang, S., Yang, F.: Research on blockchain industry news monitoring integrating semantics and emotion analysi. Modern Intell. **40**(11), 22–33 (2020). (**in Chinese**)
26. Lin, M., Wang, H., Xu, Z., Yao, Z., Huang, J.: Clustering algorithms based on correlation coefficients for probabilistic linguistic term sets. Int. J. Intell. Syst. **33**(12), 2402–2424 (2018)
27. Lu, X., Guo, Y., Chen, J., Wang, F.: Topic change point detection using a mixed Bayesian model. Data Min. Knowl. Disc. **36**(1), 146–173 (2022)
28. Rashid, J., Shah, S.M.A., Irtaza, A.: Fuzzy topic modeling approach for text mining over short text. Inf. Process. Manage. **56**(6), 102060 (2019)
29. Gao, Y., Feng, S.: A news text sentiment analysis algorithm combined with the attention mechanis. New Industrializ **10**(07), 15–18 (2020). https://doi.org/10.19335/j.cnki.2095-6649.2020.07.006. (**in Chinese**)
30. Xu, X., Tian, K.: A new method of stock index prediction based on Emotion Analysis of Financial Text. Quantitative Econ. Tech. Econ. Res. **38**(12), 124–145 (2021). https://doi.org/10.13653/j.cnki.jqte.2021.12.009. (**in Chinese**)
31. Xu, J., Wu, Y.: Research on news text classification based on the BERT-BiLSTM-CNN model. Softw. Eng. **26**(06), 11–15 (2023). https://doi.org/10.19644/j.cnki.issn2096-1472.2023.006.003. (**in Chinese**)
32. Zhai, Z., Zhang, X., Fang, F., Yao, L.: Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model. Multimed. Tools Appl. **82**(14), 20975–20988 (2023)
33. Guo, J.F., Ji, Q.: How does market concern derived from the Internet affect oil prices? Appl. Energy **112**, 1536–1543 (2013)
34. Satpathy, P., Kumar, S., Prasad, P.: Suitability of Google Trends™ for digital surveillance during ongoing COVID-19 epidemic: a case study from India. Disaster Med. Public Health Prep. **17**, e28 (2023)
35. Yu, L., Zhao, Y., Tang, L., Yang, Z.: Online big data-driven oil consumption forecasting with Google trends. Int. J. Forecast. **35**(1), 213–223 (2019)
36. Tang, X., Dong, M., Zhang, R.: Consumer confidence index prediction study based on machine learning LSTMUS model. Stat. Stud. **37**(07), 104–115 (2020). https://doi.org/10.19343/j.cnki.11-1302/c.2020.07.009. (**in Chinese**)
37. Xu, Q., Bo, Z., Jiang, C., Liu, Y.: Does Google search index really help predicting stock market volatility? Evidence from a modified mixed data sampling model on volatility. Knowl.-Based Syst. **166**, 170–185 (2019)

38. Yang, Y., Guo, J.E., Sun, S., Li, Y.: Forecasting crude oil price with a new hybrid approach and multi-source data. Eng. Appl. Artif. Intell. **101**, 104217 (2021)
39. Li, X., Liu, Y., Zhu, S.: Forecpredict implied volatility changes using Internet search attention: analysis based on artificial neural networ. Theory Pract. Syst. Eng. **43**(07), 2055–2071 (2023)
40. Antweiler, W., Frank, M.Z.: Is all that talk just noise? The information content of internet stock message boards. J. Financ. **59**(3), 1259–1294 (2004)