Journal of Computers Vol. 35 No. 4, August 2024, pp. 75-91
doi: 10.53106/199115992024083504006

STSB Model Based on STL Decomposition Algorithm and Its Application in Stock Price Prediction Studies

Chang-Sheng Zhu[1*], Yan-Bo Wang[1], Wen-Fang Feng[2], and Pei-Wen Yuan[1]

[1] School of Computer and Communication, Lanzhou University of Technology,
Lanzhou 730050, Gansu, PRC
867320505@qq.com, 1663806213@qq.com, 949067244@qq.com

[2] School of Economics and Management, Lanzhou University of Technology,
Lanzhou 730050, Gansu, PRC
1036784024@qq.com

Received 18 October 2023; Revised 20 February 2024; Accepted 14 March 2024

Abstract. Incorporating investors' stock review feature vectors in stock prediction models to improve the accuracy of stock price prediction. A stock price model STSB based on STL decomposition algorithm is proposed, which integrates the investor's stock review feature vector constructed by the BERT model and the number of reads and comments of the stock reviews as features with the stock price features extracted by the STL decomposition algorithm through the attention mechanism to make stock price prediction. Comparing with the experimental results of five models, namely, KNN, ANN, DT, SVR, and RF, on six types of stock datasets, the results show that the accuracy of the STSB model in predicting the six types of stock prices is significantly improved compared with other single benchmark models. The STSB model with the addition of STL decomposition algorithm is analyzed through experimental validation to have better prediction effect and better prediction ability for the future stock price trend.

Keywords: stock price prediction, STL, deep learning, STSB

1 Introduction

The stock market is an important part of the financial market, and it can reflect the development of the national economy to a certain extent. Predicting stock market trends in advance is important for national regulatory authorities. It helps them understand economic operations and adjust policies to keep the economy stable and healthy. Meanwhile, the stock market is also the most important part of the investment market. Compared with other industries, the stock market is characterized by both high return on investment and high risk. Effective prediction of stock prices can help investors avoid high risks when investing. Therefore, forecasting stock prices is of great importance to policy makers and investors.

Many factors affect stock prices, leading to complex, changing, and noisy patterns. This complexity makes accurate price predictions very difficult. Therefore, many models and techniques have been used to predict the closing price of stock market. In a series of recent studies, scholars have explored the effectiveness of using decision tree-based models, such as decision trees, random forests, and XGBoost, to predict stock prices. These studies have used a variety of evaluation metrics including accuracy and mean square error (MSE) with the aim of identifying the model that performs optimally in the stock price prediction task. Yifan Wu [1] et al. showed that the random forest model is considered to be the most suitable model for stock price prediction due to its excellent performance in terms of accuracy and MSE, and that in comparison to the XGBoost model, the random forest is less sensitive to the Compared to the XGBoost model, the random forest is less sensitive to the size of the input data and demonstrates better stability and reliability. In addition, the decision tree regression model was also demonstrated to have superior predictive ability compared to the linear regression model in predicting Apple's stock price in a study by Zongze Li [2] et al. This further confirms the effectiveness of the tree-based model in dealing with complex stock price prediction problems. Other studies such as Solanki Aryan [3] et al. explored a new approach to stock price prediction using Twitter data by combining Long Short-Term Memory Networks (LSTM) and Random Forest Models, and although faced with the challenges of data quality and algorithmic

[*] Corresponding Author

complexity, this hybrid model revealed the potential of social media data in predicting market dynamics. In addition, an improved random forest model based on Pearson coefficients developed by Zhengxu Yan [4] et al. made significant progress in reducing the feature selection problem and improving the prediction accuracy in big data environments, highlighting the importance of model optimization in enhancing prediction. Yuping Huang [5] et al. explored the use of hyper vector regression, random forests, and K-nearest neighbor algorithms to predict Google's share price data from 2 January 2018 to 31 March 2023 with effectiveness. By testing, analyzing and comparing the performance of these algorithms and evaluating them on the basis of Mean Square Error (MSE), the study found that the Random Forest algorithm performs the best in stock price prediction. Further using the random forest algorithm to predict the stock price trend for the next ten days, the experimental results found that the closing price will be on an upward trend, providing valuable information to investors. Random forest and decision tree models have demonstrated their advantages in stock price prediction in several studies, not only due to their high accuracy and low error rate, but also due to their high adaptability and robustness to complex data. These research results provide a strong guidance for future work in stock price prediction model selection and optimization.

Support vector regression (SVR), as an important machine learning algorithm, has received much attention for its application in improving the accuracy and stability of prediction models. Jiali Deng [6] et al. predicted stock prices by introducing the MTICA algorithm based on Tukey M estimation and the SVR model combined with AEO optimization, i.e., the MTICA-AEO-SVR model, with the It aims to improve the stability and efficiency of Fast ICA algorithm. Through the empirical analysis of SSE B shares, this shows that the MTICA-AEO-SVR model combined with AEO-optimised SVR prediction capability on the basis of denoising and independent component analysis significantly outperforms the traditional ICA-SVR model combination, signifying that prediction accuracy can be effectively improved by algorithmic optimization and combination in the field of stock price prediction. Meanwhile, the study by Yiqing Liu [7], and others explored the application of Support Vector Regression (SVR), Random Forest (RF) and Bagging-based integrated learning methods in stock futures market forecasting. By comparing with the traditional model and evaluating the model performance using R2 and MSE metrics, the results demonstrate that the Bagging-based integrated algorithm has significant advantages in enhancing the model robustness and generalization ability, and the integrated model demonstrates improvements in both R2 and MSE metrics compared to a single model, thus providing a new perspective for stock market forecasting.

K-nearest neighbor (KNN) algorithm in this field and its performance comparison with other algorithms. A study by Vineet G. Kowti [8], evaluated the effectiveness of Long Short-Term Memory Networks (LSTMs) with KNN algorithms in dealing with stock price prediction problems. By carefully collecting and preprocessing data including historical prices, trading volume and market sentiment, they built and trained the corresponding models. The performance was evaluated using metrics such as mean square error (MSE), and the results showed that the LSTM model was significantly better than the KNN model in terms of prediction accuracy. On the other hand, a study by Jiawei Pang [9] et al. compared the performance of linear regression, KNN, and decision tree algorithms in predicting Netflix stock prices. By using Mean Absolute Error (MAE), Mean Square Error (MSE) and Coefficient of Determination (R-squared) as evaluation metrics, it was found that linear regression model demonstrated optimal performance in this application scenario. Although the KNN algorithm provides effective data-driven support for stock price prediction in some cases, its performance may be affected by the nature of the data employed and the pre-processing methodology

The respective studies conducted by Xingyu Liu [10], and his team and Kocaoğlu Doğangün [11] et al. are dedicated to exploring and evaluating the efficacy and limitations of the application of machine learning models in the field of stock price prediction. Xingyu Liu [10] et al. focused on the comparison of the performance of the three models - logistic regression, KNN (K-nearest neighbor) and SVM (Support Vector Machine). A comparison of the performance of the three models in predicting stock prices, especially in the scenario of predicting the rise or fall of stock prices one day in the future based on data from the past seven days, showed that the SVM model outperforms the logistic regression and KNN models in terms of accuracy. This finding highlights the potential advantages of SVM in dealing with the stock price prediction problem and also reflects the differences in the applicability of machine learning models in different scenarios. Kocaoğlu Doğangün [11] et al. extended their research by applying XGBoost, SVM, KNN and Random Forest (RF) models to predict the stock prices of companies in different sectors of the Istanbul Stock Exchange BIST 30 index for the stock prices of companies in different sectors, which were comprehensively evaluated using assessment indicators such as MAE, MAPE, MSE and R2, with a special focus on the steel and oil sectors. The study also considers the impact of macroeconomic variables, such as the change of central bank governor, on forecasting performance, revealing the importance of macroeconomic factors in stock price forecasting and the possible negative impact of policy changes on model performance. Not only does it deepen the understanding of the application of SVM and other machine learning models in stock price forecasting, but it also highlights the multiple factors that must be considered in practice, including model selection, the setting of the data time window, and the impact of macroeconomic changes.

Bowen Ma [12], and his team and researchers such as Hota Lopamudra [13], explored the efficacy of Artificial Neural Networks (ANN) and its comparison with other prediction models such as ARIMA, Random Forest (RF), Support Vector Machines (SVMs), and Long Short-Term Memory (LSTMs), respectively, for stock price prediction. Bowen Ma [12], and others focused on analyzing the prediction accuracy of ANN vs. ARIMA for Nvidia stock price volatility during the period June 2020 to June 2021, a period when Nvidia stock price fluctuated significantly due to new product releases and global cryptocurrency price increases. By utilizing data provided by Kaggle and Yahoo Finance, the study found that ANNs had a significant advantage in capturing the significant fluctuations in Nvidia's share price, thus highlighting the potential value of artificial neural networks in dealing with complex financial data analysis. Similarly, a study by Hota Lopamudra [13] et al. further demonstrated the efficiency of ANNs in identifying and predicting stock price fluctuations by comparing the application of Random Forest, Support Vector Machines, Long and Short-Term Memory Networks, and Artificial Neural Networks in stock market prediction, especially when analyzing the performance of Random Forests and Artificial Neural Networks. In addition, the study comprehensively evaluated ANN and RF models by designing a candlestick model to show the fluctuation of stock prices over a specific period of time, aiming to identify the optimal model in stock market forecasting. The importance of artificial neural networks in the analysis of financial markets, especially in predicting stock price fluctuations, is highlighted. The ANN model stands out in these studies mainly due to its powerful non-linear modelling capabilities and its efficiency in dealing with large-scale, complex financial datasets.

It also combines the impact of online public opinion on the stock market to predict stock prices, and the integrated application of online public opinion and investor sentiment analysis in the stock price trend prediction model significantly improves the prediction performance. Researchers such as Shuaibin Zhao [14], employed a bidirectional long and short-term memory network (BiLSTM) combining empirical modal decomposition (EMD), investor sentiment analysis, and attention mechanism, and successfully improved the accuracy, precision, recall, and F1 value of stock price prediction, which was significantly better than the traditional LSTM, Attention-LSTM, SVM, and XGBoost models. In addition, Zihan Wang [15] et al. demonstrated the key role of investor sentiment index in improving the accuracy of stock price prediction by combining technical indicators and improved Snow NLP sentiment analysis model, which reduces the error indicators such as MAE, MSE, etc. Xiaoyuan Fan [16] et al. proposed a real estate stock price prediction model based on investor sentiment by comprehensive sentiment index constructed by combining Baidu search index and stock bar comment scores, and using principal component analysis (PCA) technique for data dimensionality reduction and convolutional neural network (CNN) model for stock price prediction, which provides a new perspective for understanding the impact of investor sentiment on stock price fluctuations. Yawen Yu [17] et al. analyzed the comments on the Internet through text analytics technique, using a BP neural network model combined with online public opinion and technical indicators to improve the accuracy of stock price prediction, aiming to provide scientific investment decision support for retail investors. The S_I_LSTM method introduced by Sheng-Hsiu Wu [18] et al. uses long and short-term memory networks to predict stock prices by integrating multiple data sources and investor sentiment analysis, showing the practical application of the effectiveness, outperforming traditional forecasting techniques. Together, these studies demonstrate the importance of online public opinion and investor sentiment analysis in stock price prediction, which can effectively capture and exploit market sentiment fluctuations through the combination of advanced machine learning techniques and sentiment analysis, providing new dimensions and tools for financial market analysis and prediction.

In the existing research field of stock market closing price prediction, the vast majority of methods rely on structured data analysis, mainly historical closing price information, while ignoring the potential impact of unstructured data, such as internet comments, on investor sentiment and stock market closing prices. The pre-processing and decomposition methods of stock price data that are commonly used are relatively simple and lack in-depth analyses of multi-dimensional stock price movements, resulting in limited prediction accuracy. In particular, when using stock market commentaries for prediction, most studies rely only on the textual content of the commentaries, ignoring the key factors of the readership of the commentaries and the number of comments, which have a significant impact on the accurate prediction of stock prices. In addition, in the process of extracting stock market comment vectors, simple vector extraction methods may lead to less accurate prediction results. Current prediction models usually use basic models such as Artificial Neural Network (ANN), K-nearest neighbor (KNN), Random Forest (RF) or Support Vector Regression (SVR) combined with optimization algorithms to make predictions; however, these models may face the problems of overfitting, high computational complexity and insufficient prediction accuracy. To address these problems, this study builds on existing work to collect and quantify online comment information of 178 stocks on the Shanghai Stock Exchange through crawler technology. This study adopts the Bert model to vectorize the comment content and applies the global average pooling technique, which not only reduces the risk of model overfitting, but also reduces the computational complexity, and at the same time preserves the spatial structure information of the comment feature vectors, which enhances the model's generalization ability. In addition, this study decomposes the stock price series by fusing the key features of readership and number of comments of reviews and using the seasonal and trend decomposition (STL) algorithm to subdivide the stock price series into trend, seasonal and residual components. Through the STL decomposition, this study clearly identifies the long-term trend and cyclical patterns of stock price movements, which significantly improves the accuracy of stock price forecasting. The stock price prediction is performed by constructing investors' stock review feature vectors with global average pooling as well as the number of reads and comments of the stock reviews as features to be fused with the stock price features extracted by the STL decomposition algorithm, and comparing the experimental results with five models, namely, KNN, ANN, DT, SVR, and RF, on the six types of stock datasets, the experimental results show that the STSB model compares favourably with the other single benchmark The experimental results show that the STSB model has significantly improved the prediction accuracy of the six types of stock prices compared to other single benchmarks. It can be seen that the STSB model has significantly improved in terms of error control and prediction accuracy, and the introduction of the STL decomposition algorithm can improve the accuracy of stock price prediction and the generalization ability of the model prediction. This paper provides a novel method and idea for stock market closing price prediction based on online public opinion.

2 Relevant Theory and Methodology

2.1 Principles of STL Decomposition Algorithm

STL (Seasonal and Trend decomposition using Loess) is a time series decomposition method, the basic principle of which is to smooth the time series data by robust locally weighted regression (Loess) and decompose the stock price time series Y_t into three main components, including the trend quantity T_t, the seasonal quantity S_t and the Residual quantity R_t, as shown in the following equation.

$$Y_t = T_t + S_t + R_t,\ t = 1, 2, ..., N. \tag{1}$$

The decomposition process of STL consists of two main parts: the outer loop and the inner loop. The main task of the outer loop is to adjust the robustness weights, while the inner loop is mainly responsible for the decomposition calculation of the two components of trend and seasonal quantities. The inner loop is divided into the following 6 steps.

Step 1: De-trending: In this step the trending quantity $T_t^{(k)}$ is removed, i.e. the time series is subtracted from the trending quantity of the previous iteration, where in the initial case $T_t^{(0)} = 0$.

Step 2: Loess regression: each subsequence is processed by LOESS regression, which requires extending one loop cycle before and after, with a smoothing parameter of n_s, and the result of smoothing is denoted as $C_t^{(k+1)}$.

Step 3: Low-pass filtering: The smoothing result $C_t^{(k+1)}$ in step 2 is subjected to a low-pass filtering process, firstly, a sliding average of lengths n_p, n_p, 3is carried out, and then a Loess regression with parameter n_1 is carried out to obtain the sequence $L_t^{(k+1)}$ of length N, which is equivalent to a low-pass filtering of the cyclic subsequence.

Step 4: Seasonal term: Remove the low-pass filtered portion of the smoothed periodic subsequence to get the seasonal term $S_t^{(k+1)} = C_t^{(k+1)} - L_t^{(k+1)}$.

Step 5: De-seasonalization: the time series Y_t is subtracted from the seasonal term $S_t^{(k+1)}$ to obtain the de-seasonalized series.

Step 6: Trend smoothing: Perform Loess regression with parameter n_t on the sequence obtained in step 5 to obtain the trend quantity $T_t^{(k+1)}$, and then determine whether it has converged. If converged, output the final result; otherwise return to step 1.

Where $T_t^{(k)}$ and $S_t^{(k)}$ denote the trend and seasonal quantities at the end of the k-1st iteration in the inner loop, with $T_t^{(k)} = 0$ at the initial moment; n_s, n_1, and n_t are the Loess smoothing parameters in Steps 2, 3, and 6, respectively; and n_p is the number of cycle samples. The number of iterations of the outer loop is denoted by n_o . When n_i is large enough, the trend and period components have converged at the end of the inner loop, and n_o can be set to 0 if there are no obvious outliers in the time series data.

2.2 Extraction and Global Average Pooling of Stock Review Vectors Using Bert

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art pre-trained linguistic representation model that encodes textual data into high-dimensional vector forms. In the context of stock market comment analysis, the semantic information of text can be effectively obtained by extracting feature vectors using BERT and through global average pooling operation. The specific process is as follows: firstly, the dataset of stock market commentaries is collected and prepared, which contains various stock market related commentaries texts. Next, a pre-trained BERT model is loaded using Hugging Face’s Transformers library, and the version of the model that best matches the particular analysis task is selected. Subsequently, each comment text in the dataset was subjected to a disambiguation process to convert it into a format that the BERT model could understand, which included adding special identifiers (e.g., [CLS] and [SEP]) and converting the text to the appropriate index. Afterwards, this processed text data is fed into the BERT model, which generates a hidden state vector for each word, which together form a high-dimensional representation of the input text. In order to extract useful features from these representations, a global average is computed over the dimensions of each hidden state vector to produce a single vector representing the semantic content of the entire comment text. This globally averaged pooled vector can then be used as a feature vector for stock market reviews for subsequent data analysis and modelling work. In summary, the steps involved in extracting stock market comment vectors and performing global average pooling using the BERT model include: data preparation, loading the pre-trained BERT model, text segmentation, model input processing, and global average pooling to obtain the final stock comment vectors.

3 Building the STSB Model Prediction Framework

With the increasing popularity of quantitative investment concepts and the successful application of deep learning models in the financial field, more and more scholars have started to use deep learning techniques for stock price prediction. This trend aims to provide more efficient and scientific decision-making tools for a wide range of investors to help them better understand the trend of stock prices in the stock market and make timely and reliable predictions of asset holdings in anticipation of unknown and uncertain events. This is an attractive option for investors. The STSB prediction model proposed in this paper mainly consists of a feature extraction module, a stock price decomposition processing module, a social engagement metrics normalization module, an embedding module for embedding, and a feature fusion module. The specific steps for constructing the STSB model in this paper are as follows.

(1) Crawl the stock reviews through Oriental Wealth Network and integrate the stock reviews, then extract the feature vectors by BERT model for each stock review, and pool the global average of the extracted feature vectors to reduce the dimensionality.

(2) By downloading the historical stock price data based on the Tushare financial interface package in Python, and then decomposing the stock price time series using the STL decomposition algorithm, and then normalizing the standard deviation of the decomposed seasonality, trend and residuals. This helps to understand the seasonality and trend components of the stock price movement, and then normalize the standard deviation of the seasonality and trend to ensure that the changes are comparable across different data ranges.

(3) In the stock reviews crawled by Oriental Wealth, each stock review contains the stock review, the reviewer, the readership of the review and the number of reviews. The readership and the number of reviews of each stock are normalized by standard deviation to ensure that the values of these two features have similar scales. Standard deviation normalization is done by subtracting the mean from each value and dividing by the standard deviation. The normalized values were included as additional features in the STSB share price prediction model to increase

the source of information for the model and to improve the accuracy of the prediction of share price movements.

(4) In order to better handle the data of multiple stocks so that the unique features of each stock can be learnt better, the embedding method is used. Embedding each stock is a technique that maps the stock into a low dimensional vector space. Instead of the original high dimensional data, each stock can be represented as a vector of values. This embedding vector contains important features and information about the stock that can allow the model to better capture similarities and differences between different stocks.

(5) The feature extraction module is used to extract feature vectors from the stock reviews, the stock price decomposition module decomposes the stock price data into trend, seasonality and noise components to better understand the fluctuation of the stock price, the social engagement metrics normalization module normalizes the number of reads and comments in the stock reviews as an indicator of attention, and the embedding module embeds the textual data into numerical representations to be fused to the features, and then fuses the four modules into the features, and then inputs the fused data into the Transformer model.

The STSB prediction process, as shown in Fig. 1.

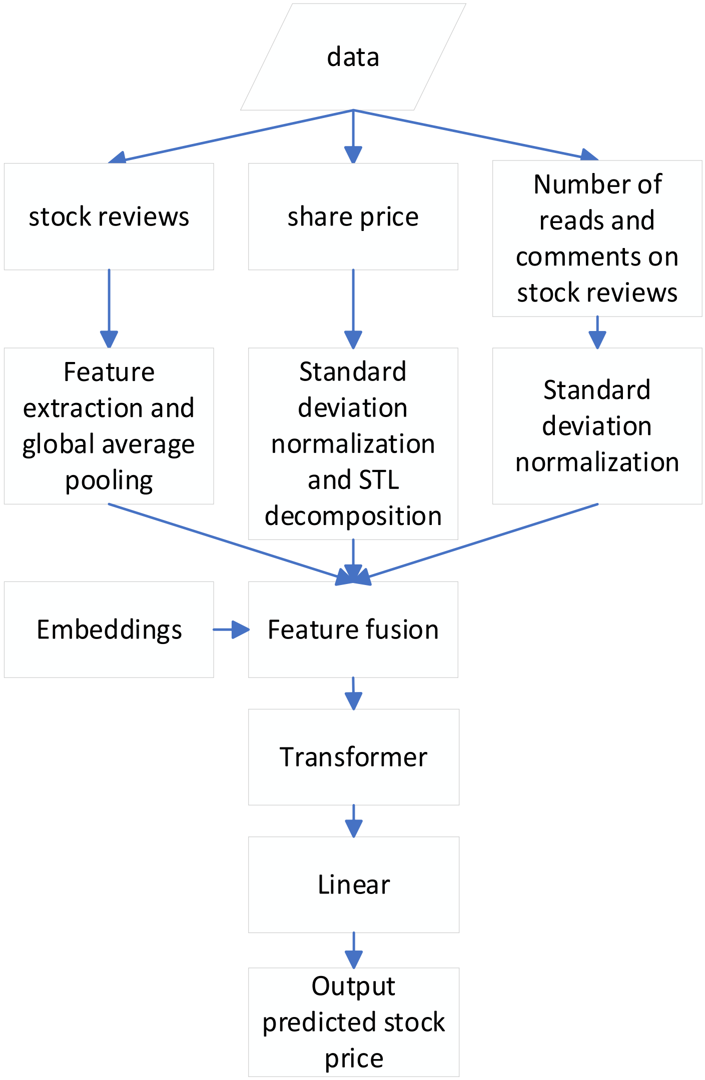

data
stock reviews
share price
Number of reads and comments on stock reviews
Feature extraction and global average pooling
Standard deviation normalization and STL decomposition
Standard deviation normalization
Embeddings
Feature fusion
Transformer
Linear
Output predicted stock price

Fig. 1. STSB prediction process

4 Experimental

4.1 Data Sources

This paper takes SSE177 stock index as the research sample, and the sample period is from January 3, 2017 to December 29, 2017. It is divided into two parts: the first part is the posting information of the Oriental Wealth Network (OWN) captured by the python web crawler program; the second part is the historical price data of stocks based on the Tushare financial interface package downloaded from Python to ensure the accuracy of the data source.

The stock text information such as posting title, number of reads, number of comments, author, date of publication is obtained by Python spyder web crawler program and stored in local disk in .xlsx format. In order to observe the prediction effect of different prediction methods on stock indices, this paper takes the last 21 days of the overall dataset of each index as the test dataset, a total of 223 sets of data from January 3rd to November 30th, 2017 as the training set, and a total of 21 sets of data from December 01st to December 29th, 2017 as the test set. In order to eliminate the magnitude gap between the data, the data were normalized using Z-score with the following equation.

$$x^* = \frac{x - \overline{x}}{\partial}. \tag{2}$$

Where, x is the original data, $\overline{x}$ and ∂ represent the mean and standard deviation of the original data, respectively, and x^* is the standardized value.

4.2 Experimental Setup

In terms of model parameter settings, the model parameter settings are finally determined by conducting several experiments, and the model training in this paper adopts the Adam optimizer with a learning rate of $1e-4$. The loss function of the prediction is MAE (Mean Absolute Error) MAE has some advantages over other regression loss functions (such as the mean square error MSE), such as insensitivity to outliers, which better reflects the absolute difference between the predicted value and the actual value of the stock price, and the Batch-size and Epochs are set to 256 and 100, respectively. In order to further verify the prediction effect of the STSB model, the model is trained using the Adam optimizer with a learning rate of $1e-4$. STSB model's prediction effect, five stock price prediction models are selected for comparison experiments, namely KNN, ANN, DT, SVR and RF.

4.3 Evaluation Indicators

In order to assess the effectiveness of the model in predicting stock prices, therefore four evaluation metrics are used which are Root Mean Square Error (RMSE) is an indicator of the root mean square error between the predicted values and the actual observed values, which penalizes large errors more heavily. Mean Absolute Error MAE (Mean Absolute Error) is a measure of the mean absolute error between predicted and actual observed values. Mean Absolute Percentage Error MAPE (Mean Absolute Percentage Error, MAPE) is an indicator of the average absolute percentage error between the predicted values and the actual observations, which expresses the magnitude of the error in percentage form. Three evaluation indexes are used to evaluate the prediction performance of the model, and the formulas for each evaluation index are as follows.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}, \tag{3}$$

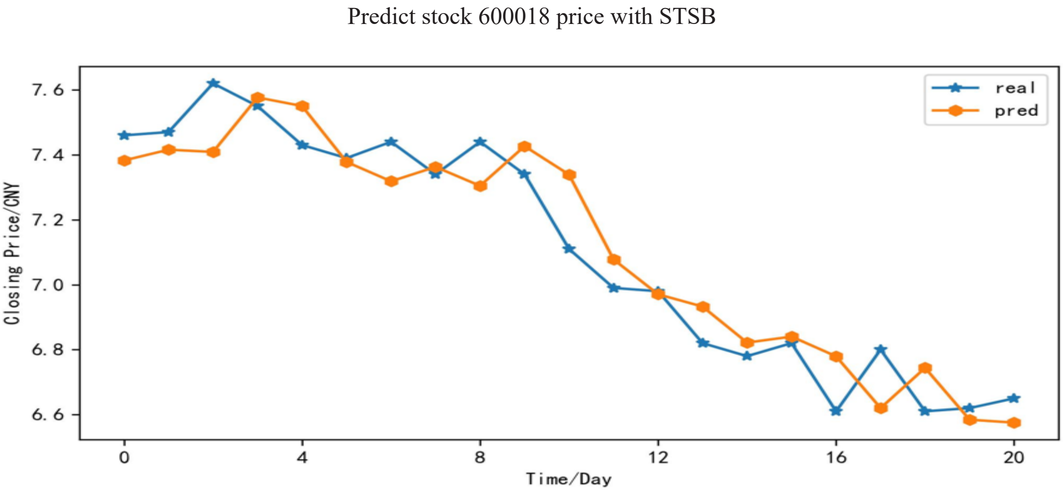
$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right|, \tag{4}$$

$$MAPE = \frac{100}{m}\sum_{i=1}^{m}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right|. \quad \mathbf{(5)}$$

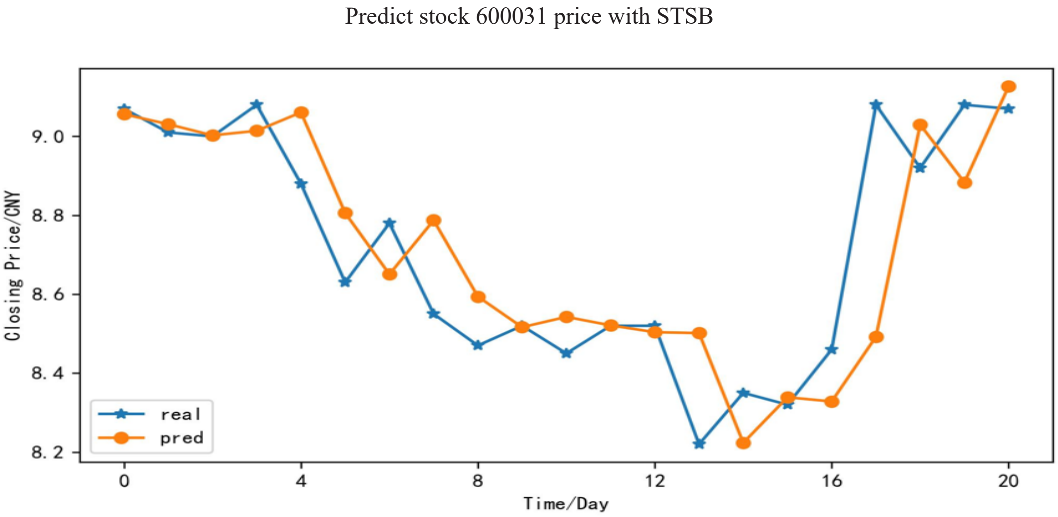
Where m is the sample size, y_i is the actual observed value, $\hat{y}_i$ is the corresponding predicted value, and $\overline{y}_i$ is the mean value of the actual observed value. RMSE, MAE and MAPE are used to measure the deviation between the actual observed values and the corresponding predicted values, and their values are in the range of [0, +∞), and the closer the value is to 0, the better the prediction effect of the model is.

4.4 Analysis of Results

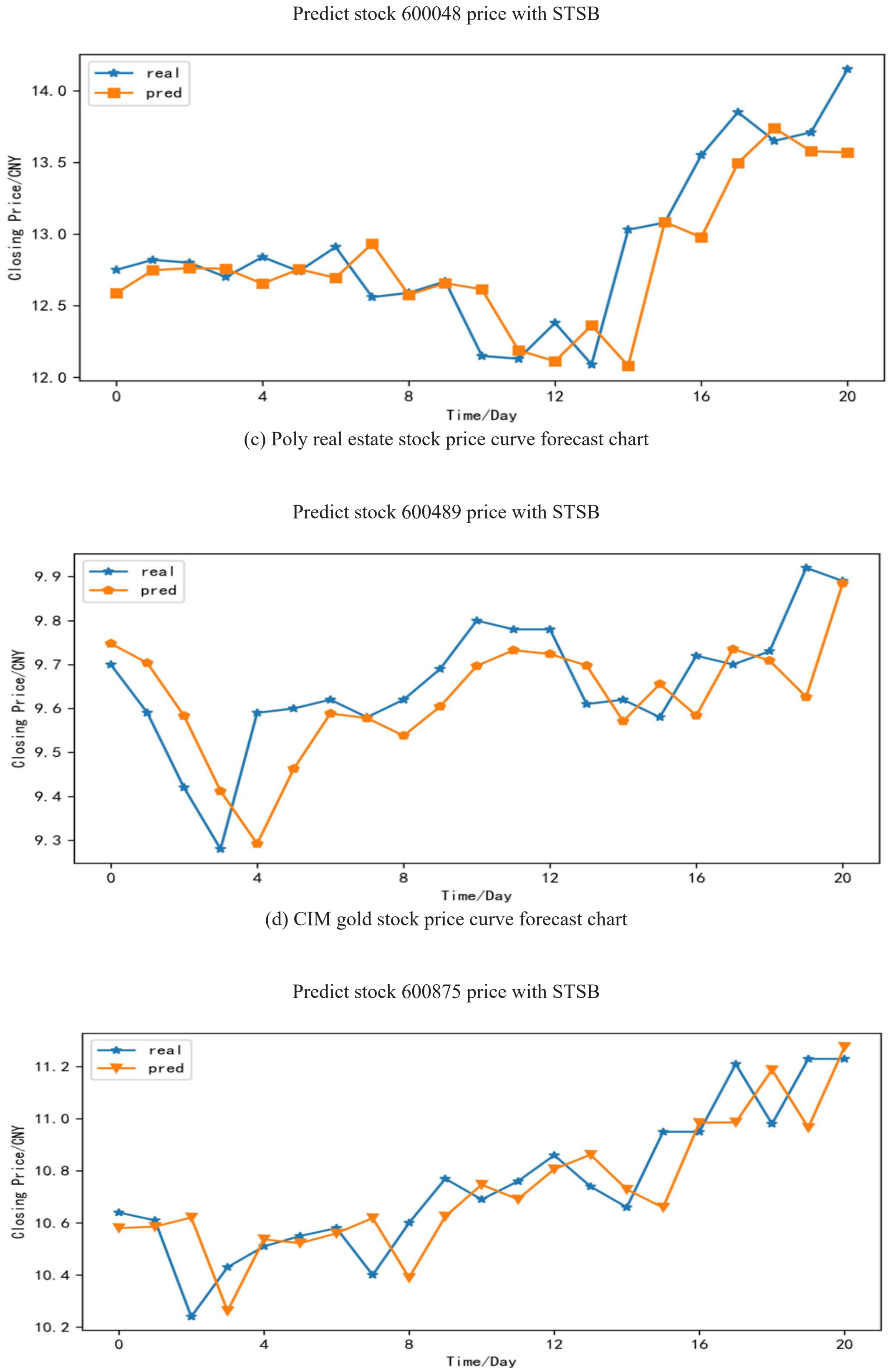
For the effect of STSB model on stock price prediction, as well as its fitting ability, the stock price prediction curves of six stocks, namely, 600018 SIPG, 600031 Sany Heavy Industry, 600048 Poly Real Estate, 600489 Zhong jin Gold, 600875 Dong fang Electric, and 601555 Soochow Securities, which are representative of the stock index of SSE 177, are specially intercepted from SSE 177. The test set is the daily frequency data of 21 days from December 01 to December 29, 2017, and the prediction label is the closing price of the next day. As shown in Fig. 2.



(a) Forecast chart of the share price curve of SIPG



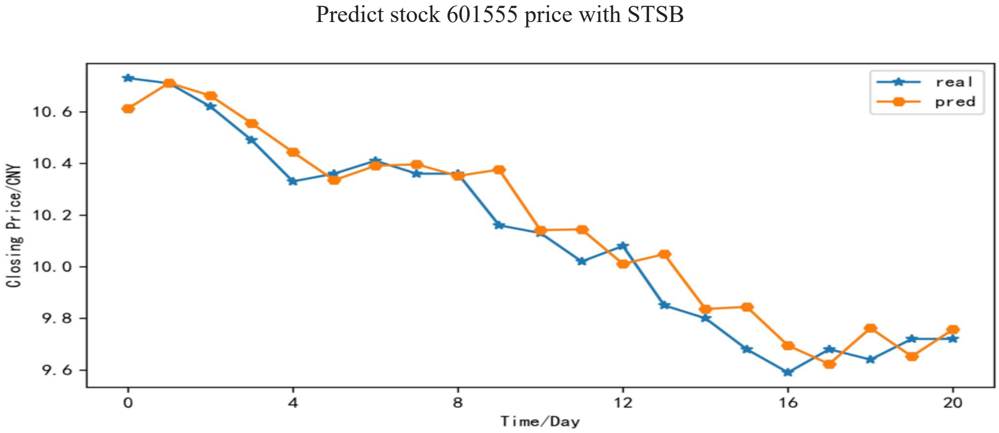
(b) Forecast chart of Trinity's share price curve



(c) Poly real estate stock price curve forecast chart

(d) CIM gold stock price curve forecast chart

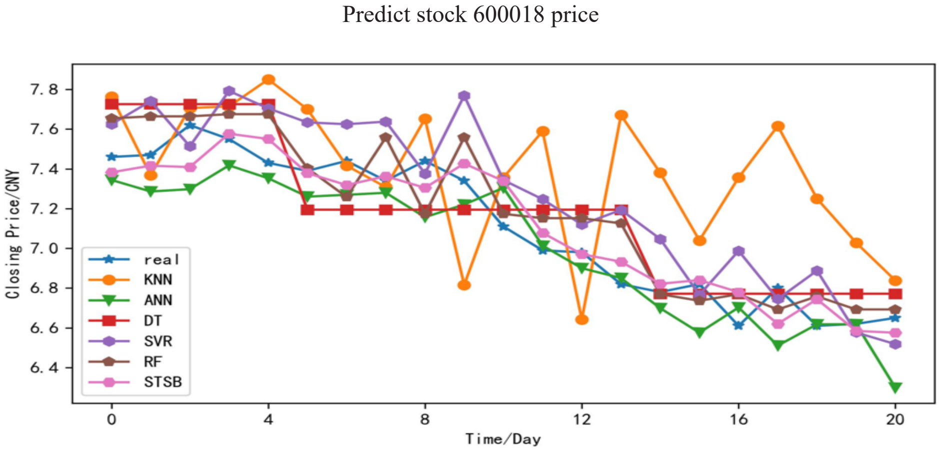
(e) Eastern electric stock price curve forecast chart



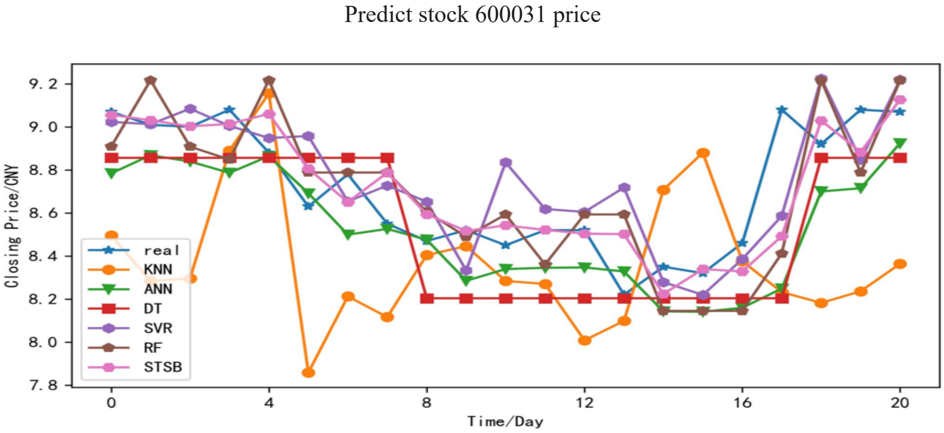
(f) Soochow securities stock price curve forecast chart

Fig. 2. STSB share price forecast curve

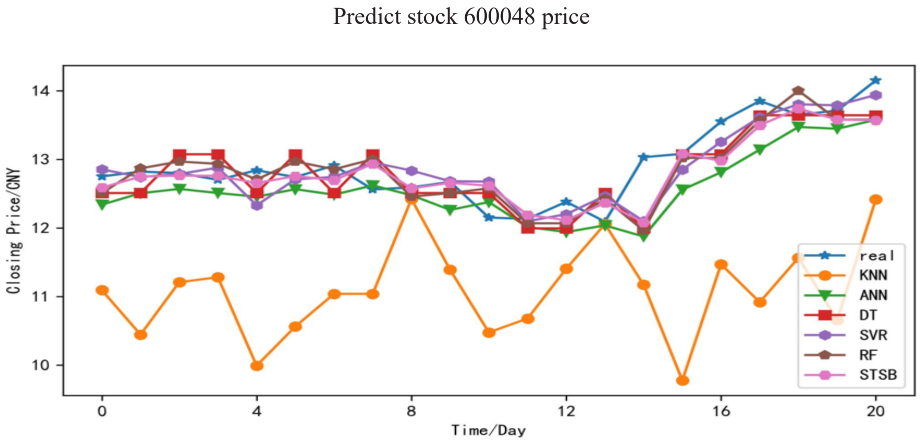
In order to verify the superiority of the model STSB, choose to compare with KNN, ANN, DT, SVR, RF model, the model are selected to have the best performance of the parameters, the test set is also selected from the test set for the test set for a total of 21 days of daily frequency data from December 01, 2017 to December 29, 2017 to the next day's closing price as the prediction label, the prediction results and the actual observed values are compared, and the same selection of the above six stocks as the experimental subjects, as shown in Fig. 3.



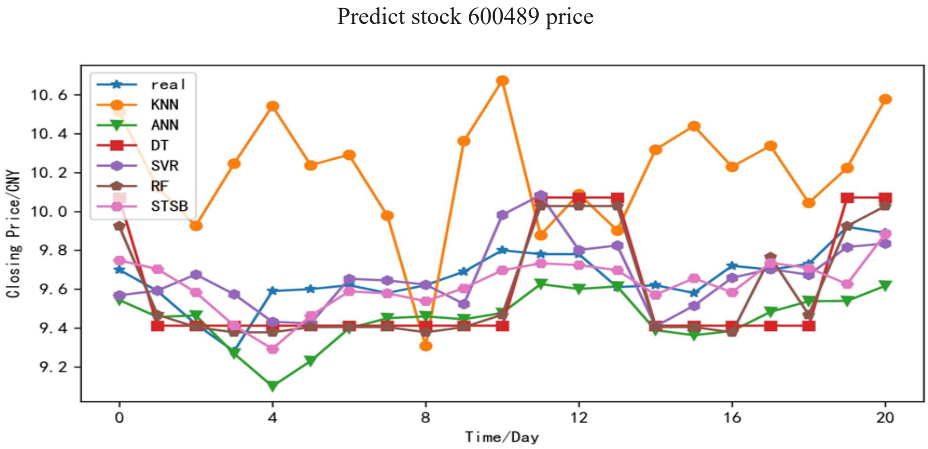
(a) Comparison of SIPG forecast model results



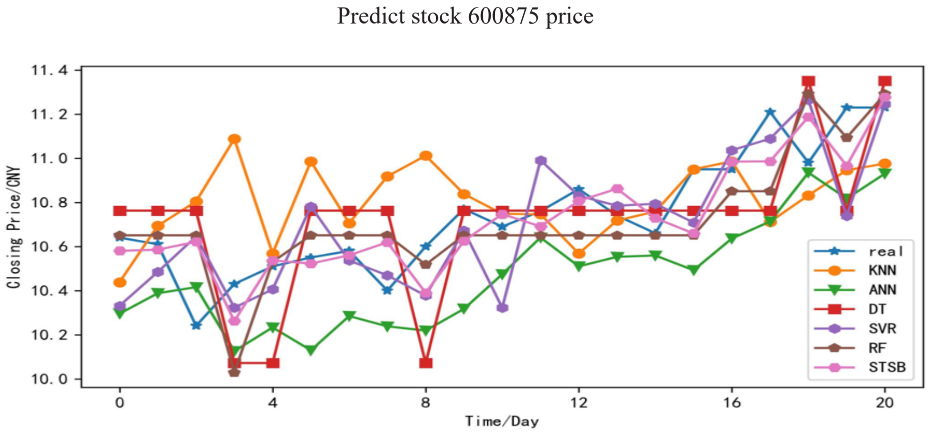
(b) Comparison chart of Sany's prediction model results



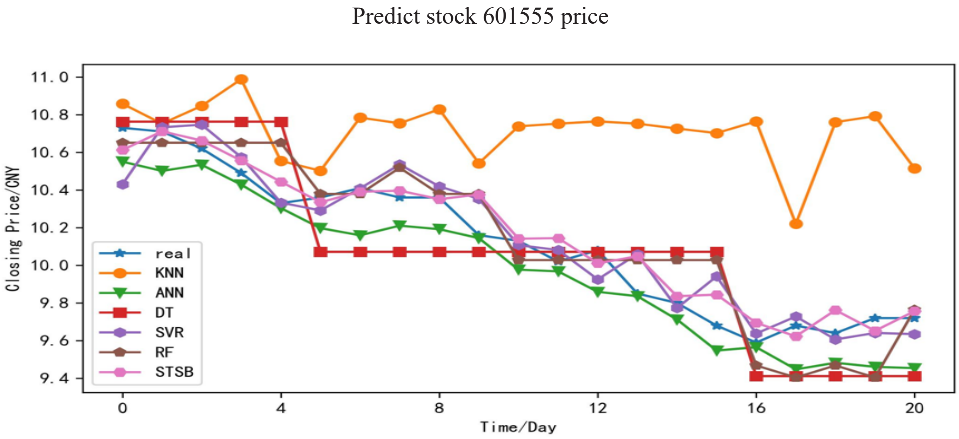
(c) Comparison of Poly Real Estate Forecast Model Results



(d) Comparison Chart of CIM Gold Forecasting Model Results



(e) Comparison of the results of the forecast model of Dongfang Electric



(f) Comparison chart of forecast model results of soochow securities

Fig. 3. Comparison of predictive model results

Comparison of prediction results shows that the new STSB model has good prediction effect on the stock price trend, for example, the stock price of Shanghai Harbor Group has fallen sharply at the high level, and the downward trend is obvious. Zhongjin Gold's stock price fluctuates in the range, and the amplitude fluctuates sharply. Poly real estate stock price low quickly climbed, the upward trend is obvious. Dongfang Electric stock price low shock climb, climbing trend is obvious. Sany Heavy Industries stock price high shock downward, swing down, and then quickly climbed rapidly, the upward trend is obvious.

In the existing research in the field of stock market closing price prediction, most of the methods mainly rely on the analysis of structured data, especially the information of historical closing prices. Since most of the studies on stock price prediction are based on simple preprocessing of the original historical stock price data, and cannot analyze the original historical stock price data in a multifaceted and comprehensive way, this study uses the STL decomposition algorithm to analyze and process the original historical stock price data in depth from a multifaceted and comprehensive perspective, in order to achieve the improvement of the accuracy of predicting the stock price. At present, there are still some studies that use social media text data, i.e., stock reviews, and historical stock price data to predict stock prices, which can predict stock prices, but the accuracy of stock price prediction is not high. This study not only fuses the social media text data, i.e., stock reviews, with the original historical stock price data, but also performs global average pooling on the extracted feature vectors of the stock reviews to solve the risk of overfitting the model, and reduces the computational complexity, retains the important feature information, and improves the accuracy of the stock price prediction, which is also assisted by the fusion of the standard deviation of the number of reads and comments on the stock reviews, which is an important feature. KNN has high computational cost, easy overfitting and poor adaptability to new trends in predicting stock prices, which leads to the low accuracy of KNN in predicting stock prices. The global average pooling module in the STSB model reduces the problems of high computational cost and easy overfitting, and thus improves the accuracy of stock price prediction. ANN relies on a large amount of historical data for training, which leads to the inability to accurately predict the uncertainty and dynamic changes in the future market. In addition, the complexity of ANNs can lead to overfitting, which results in poor accuracy of ANNs in predicting stock prices. The global average pooling module in the STSB model reduces the problem of easy overfitting, and the STL decomposition algorithm module improves the model's ability to generalize the model to predict stock prices, which is a good predictor of stock prices of different industries. The DT can easily over-simplify the complexity of the market dynamics in the prediction of stock prices, which is prone to overfitting, especially in the case of the DT. The global average pooling module in the STSB model solves the problem of overfitting, and the SVR has insufficient adaptability and predictive ability to predict stock prices in non-linear market data, especially when dealing with highly volatile and noisy financial time series data. cause the model to preferentially learn and refine generalized feature representations, which enhances the model's adaptability and robustness when encountering novel and unobserved data as compared to the reduced feature dependency of only targeting samples from a specific training set, thus improving the prediction accuracy and generalization ability for unfamiliar datasets. RF is unable to effectively deal with the effects of external factors such as non-linearity and market sentiment when predicting stock prices. For example, it is unable to capture the impact of stock market prices on complex macroeconomic indicators, breaking news events and investor sentiment fluctuations, which leads to the low accuracy of RF in predicting stock prices. The global average pooling module in the STSB model extracts more generalized features, which reduces the model's dependence on a specific training dataset, improves the model's adaptability to unseen data and enhances its generalization performance. In summary, KNN, ANN, DT, SVR, and RF have high computational cost, easy overfitting, and insufficient adaptability and predictive ability to non-linear market data in predicting stock prices, resulting in the inability to accurately predict the uncertainty and dynamic changes in the future market, which leads to the low accuracy of ANN in predicting stock prices. The STL decomposition algorithm module and global average pooling in the STSB model can solve the problems of high computational cost, easy overfitting, and insufficient adaptability and predictive ability to nonlinear market data, and poor adaptability to new trends, thus improving the accuracy of stock price prediction. And by integrating the important features of readership and comment counts of stock reviews after standard deviation normalization to assist in stock price prediction, the accuracy of stock price prediction can be better improved.

In addition, from the six model fitting effect, it can be intuitively seen that the STSB model has the best fitting effect. Calculating the mean value of the prediction error of each prediction model, and from the perspective of the three evaluation indexes, the mean value of the error of STSB has reached the minimum value. As shown in Table 1.

Table 1. Comparison of experimental predictors for each model

Model	Evaluation indicators	Shanghai Harbor Group	Sany Heavy Industry	Poly Real Estate	Centrum Gold (Chinese gold company)	Dong fang Electric	Soochow Securities (PRC stock exchange)
KNN	RMSE	0.4488	0.5288	0.4652	0.6218	0.306	0.6845
	MAE	0.3729	0.4558	0.3796	0.5737	0.23	0.593
	MAPE	0.0537	0.0517	0.1392	0.0595	0.0216	0.0598
ANN	RMSE	0.1755	0.267	0.4505	0.244	0.3128	0.1612
	MAE	0.1425	0.2058	0.3671	0.2128	0.2874	0.1385
	MAPE	0.0199	0.0234	0.0281	0.022	0.0267	0.0137
DT	RMSE	0.1945	0.2839	0.3862	0.257	0.2919	0.2474
	MAE	0.1731	0.2259	0.3162	0.237	0.2366	0.2152
	MAPE	0.0242	0.0258	0.0246	0.0245	0.0221	0.0214
SVR	RMSE	0.2401	0.2279	0.3195	0.1543	0.2215	0.1291
	MAE	0.2137	0.1797	0.2673	0.1221	0.1793	0.1083
	MAPE	0.0301	0.0207	0.0205	0.0127	0.0167	0.0157
RF	RMSE	0.1659	0.2535	0.3665	0.2257	0.2006	0.1764
	MAE	0.1441	0.2118	**0.2373**	0.2016	0.154	0.1401
	MAPE	0.0202	0.0242	**0.0185**	0.0208	0.0184	0.014
STSB	RMSE	**0.117**	**0.1981**	**0.2956**	**0.1123**	**0.1701**	**0.0899**
	MAE	**0.0948**	**0.1432**	0.2414	**0.081**	**0.1398**	**0.074**
	MAPE	**0.0135**	**0.0164**	0.0186	0.0084	0.013	0.0076

Fig. 4 shows the relative improvement percentages of the STSB model compared with the other five models in different indexes. The STSB model has obvious improvement in all evaluation indexes compared with the KNN, ANN, DT, SVR and RF models. Specifically, the RMSE and MAE of the STSB model have improved by more than ten percent relative to those of the five models. Furthermore, the MAPE of the STSB model has increased by over twenty percent compared to the MAPEs of the other models. Additionally, the STSB model surpasses the other models in forecasting the closing stock prices of six different industries, indicating superior predictive performance and a more advantageous position in stock index forecasting. This superiority further corroborates the accuracy and adaptability of the STSB model, as depicted in Fig. 4.

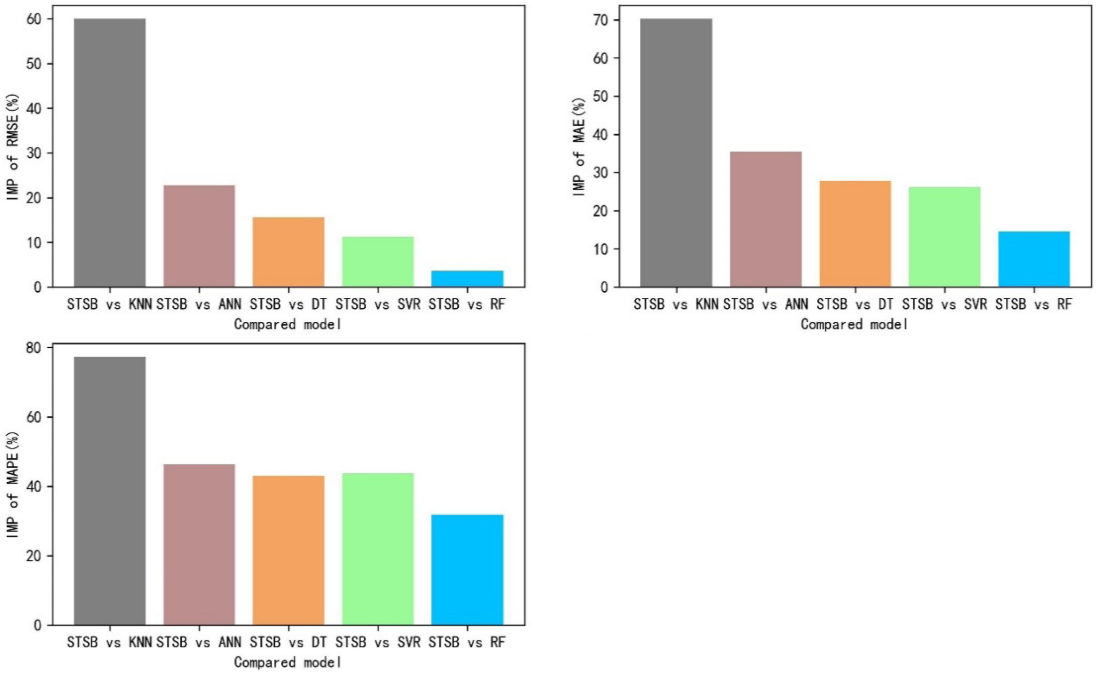


Fig. 4. Comparison of experimental predictors of STSB model with other models

Fig. 5 shows the scatter plots of real and predicted values of KNN, ANN, DT, SVR, RF, and STSB stock prices on the six models. It can be seen that the scatter plot of ANN model is more dispersed, and the slope of the fitted straight line is only 0.9564, while the slopes of the fitted straight lines of KNN, DT, SVR, and RF models are not much different from each other, and they are 0.9602, 0.966, 0.9652, 0.9673 respectively. Among the six models, the scatter plot of STSB is clustered near the ideal fitting straight line, and the slope of its straight line is the largest. The slope of the straight line fit is the largest. As shown in Fig. 5.

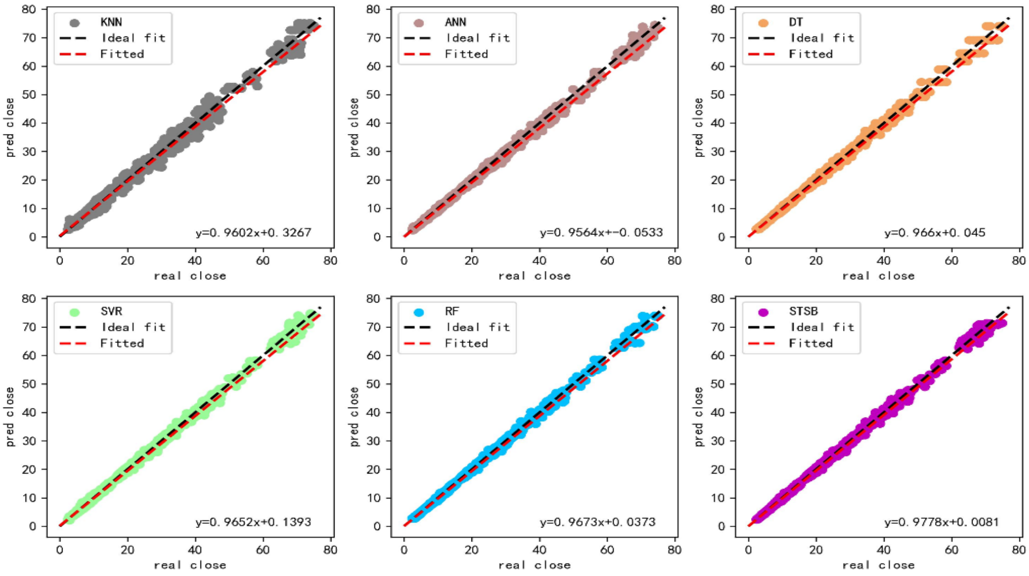


Fig. 5. Scatterplot of real and predicted values of stock price on six models KNN, ANN, DT, SVR, RF, STSB

Predictive residual plots can help to identify non-systematic components in stock price movements to better understand the causes and influences of stock price fluctuations, and the predicted residual values can provide information to modify the valuation model to improve the accuracy of stock valuation. From the figure, it can be seen that the residual value of KNN model is the largest of the six models, which is 1.042, indicating that the residuals of the model are very unstable, and the predicted stock value is prone to extreme values, thus affecting the accuracy of stock price prediction, and the residuals of ANN, DT, SVR, RF are 0.48, 0.50, 0.471, 0.438 respectively, which are closer to each other, and the residuals of KNN model are 0.471 and 0.438 respectively, which are closer to the residuals of KNN model, and the residuals of KNN model are 0.48, 0.50, 0.471 and 0.438, which are closer to each other than KNN model. KNN model has some improvement, but the residual value of STSB model is 0.42, which indicates that the residual of STSB model is more stable, compared with the other five models, the stability of the residual is significantly improved, and it can more accurately predict the stock price and increase the accuracy of stock price prediction. As shown in Fig. 6.

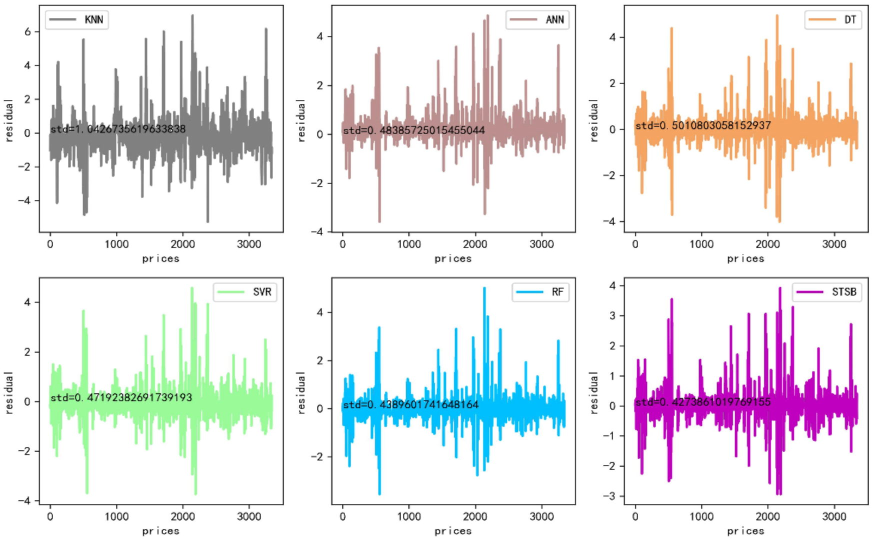


Fig. 6. The prediction residuals of the six models KNN, ANN, DT, SVR, RF, STSB

4.5 Ablation Experiment

In order to verify the prediction effect of the proposed STSB model and the validity and necessity of each functional module, a variant model is obtained by removing different modules in the module, and an ablation experiment is designed and evaluated according to the calculation of evaluation indexes.

The STL decomposition algorithm ablation experiment is conducted by controlling the other parts of the model to remain unchanged and deleting only the part of the STL decomposition algorithm module in the model, and comparing the complete STSB model with the STSB variant model with the STL decomposition algorithm module deleted. The results of the comparison between the complete STSB model with the STL decomposition algorithm added and the STSB variant model with the STL decomposition algorithm removed are shown in Table 2.

As shown in Table 2, by comparing the results of RMSE, MAE and MAPE, the complete STSB model with STL decomposition algorithm outperforms the STSB variant model with STL decomposition algorithm removed, and the RMSE, MAE and MAPE are reduced by an average of 7.18%, 9.79% and 12.84%, respectively, which verifies the validity of the addition of STL decomposition algorithm. After analyzing the basic components of the stock price, the STL decomposition algorithm can model the trend and seasonal components separately, thus improving the overall forecasting accuracy. The STL decomposition algorithm is highly flexible, and can be adapted to a variety of time series data characteristics, including those with non-linear trends and complex seasonal patterns. The STL decomposition algorithm can effectively separate seasonality, trend and residuals, which can help to understand and forecast the stock price more accurately. It helps to understand and predict more accurately the cyclical fluctuations, long-term trends and abnormal fluctuations of stock prices. It also shows that the STSB model with the inclusion of STL decomposition algorithm has better forecasting effect, which further demonstrates the superiority of the inclusion of STL decomposition algorithm with better forecasting ability for future stock price trends.

Table 2. STL decomposition algorithm ablation experiment results

Model	Evaluation indicators	Shanghai Harbor Group	Sany Heavy Industry	Poly Real Estate	Centrum Gold (Chinese gold company)	Dongfang Electric	Soochow Securities (PRC stock exchange)
STSB	RMSE	**0.117**	**0.1981**	**0.2956**	**0.1123**	**0.1701**	**0.0899**
	MAE	**0.0948**	**0.1432**	**0.2414**	**0.081**	**0.1398**	**0.074**
	MAPE	**0.0135**	**0.0164**	**0.0186**	**0.0084**	**0.013**	**0.0076**
STSB variant	RMSE	0.1251	0.2087	0.3263	0.1219	0.1806	0.0964
	MAE	0.1145	0.1585	0.2534	0.097	0.1497	0.0851
	MAPE	0.0168	0.0179	0.0197	0.0095	0.0161	0.0087

5 Conclusion

In this study, the paper construct a stock price prediction model by integrating the STL decomposition algorithm with BERT and Transformer; firstly, the paper extract the feature vectors of stock reviews through the Bert model, and pool the global average of the feature vectors of stock reviews; and normalize the standard deviation of the number of reads and comments of each stock review, and then the paper use the standard deviation-normalized number of reads and comments of each review as a feature to make the stock price prediction more accurate. The standard deviation normalized readership and comment counts of the reviews are used as a feature to make the stock price prediction more accurate. The standard deviation normalized stock prices are processed and then the STL decomposition algorithm is applied to the standard deviation normalized stock prices so that the stock prices can be predicted more accurately; and the embedding of each stock is performed so that the unique features of each stock can be learnt better. Finally, the dataset is divided into training set and test set and input into the STSB model as samples for correlation prediction, and the prediction results are significantly improved.

The STSB model predicts the stock prices of 177 stock indices of SSE, and the following conclusions are drawn: (1) When the constructed STSB model predicts the stock prices of six different industries, it has a good prediction ability, which reflects the generalization ability of the newly built model in stock price prediction. (2) By comparing the STSB model with KNN, ANN, DT, SVR and RF models, it can be seen that the STSB model has significantly improved in terms of error control and prediction accuracy, which can help investors to find more investment opportunities and improve the return on investment. (3) In the stock market, it can be found that there is a strong correlation between the stock price and the stock rating, and the refined data processing of the stock price can better predict the stock price, so the STL decomposition algorithm is introduced, and the generalization ability and accuracy of the model prediction after the introduction of the STL decomposition algorithm is verified through empirical analysis. In the future research work, when predicting the stock price, by extracting the stock review feature vector, there is anisotropy in the stock review feature vector, and the anisotropy will have a negative impact on the performance of the model, resulting in a decrease in the accuracy of the stock price prediction. In the future work, it can be considered to eliminate the effect of anisotropy, so as to improve the accuracy of the stock price prediction.

6 Acknowledgement

National Natural Science Foundation of China Regional Science Fund “Research on the formation mechanism and monitoring system of asset price bubbles based on the perspective of financial leverage” (Grant Nos: 72161026).

References

[1] Y.-F. Wu, Stock Price Prediction Based on Simple Decision Tree Random Forest and XGBoost, BCP Business & Management 38(2)(2023) 3383-3388.

[2] Z.-Z. Li, Comparison of Decision Tree Regression with Linear Regression Based on Prediction of Apple Stock Price, Advances in Economics, Management and Political Sciences 45(2)(2023) 62-69.

[3] S. Aryan, S. Singla, H. Petwal, An Improved Machine Learning Algorithm for Stock Price Prediction, in: Proc. 2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon), 2023.

[4] Z.-X. Yan, C. Qing, G. Song, Random Forest Model Stock Price Prediction Based on Pearson Feature Selection, Computer Engineering and Applications 57(15)(2021) 286-296.

[5] Y.-P. Huang, Research on the Google Stock Price Prediction Based on SVR, Random Forest, and KNN Models, Highlights in Business, Economics and Management 24(2)(2024) 1054-1058

[6] J.-L. Deng, F.-Q. Zhao, X.-X. Wang, MTICA-AEO-SVR model for stock price forecasting, Computer Engineering and Applications 58(8)(2022) 257-263.

[7] Y.-Q. Liu, A. Ayitelieke, J.-Y. Yu, Short-Term Stock Price Prediction Algorithm Construction Based on Integrated Learning of SVR and RF with Bagging, Highlights in Science, Engineering and Technology 22(2022) 8-15.

[8] V.G. Kowti, Stock Price Prediction using LSTM and KNN Algorithms, International Journal for Research in Applied Science and Engineering Technology 11(IX)(2023) 1591-1595.

[9] J.-W. Pang, Netflix Stock Price Prediction: A Comparison of Linear Regression, k-Nearest Neighbor, and Decision Tree Methods, Highlights in Business, Economics and Management 24(2)(2024) 569-574.

[10] X.-Y. Liu, Y.-Z. Ni, B.-T. Yang, Stock Price Prediction of Apple Based on SVM and KNN, BCP Business & Management 34(5)(2022) 873-878

[11] K. Doğangün, K. Turgut, M.Z. Konyar, Sector-Based Stock Price Prediction with Machine Learning Models, Sakarya University Journal of Computer and Information Sciences 5(3)(2022) 415-426.

[12] B.-W. Ma, Y.-C. Yang, J.-M. Zhang, K.-L. Zhang, A Comparison of Stock Price Prediction with ANN and ARIMA, BCP Business & Management 38(2)(2023) 392-399.

[13] H. Lopamudra, P.K. Dash, Comparative Analysis of Stock Price Prediction by ANN and RF Model, Computational Intelligence and Machine Learning 5(1)(2021) 2582-7464.

[14] S.-B. Zhao, X.-D. Lin, X.-J. Weng, Attention-BiLSTM stock price trend prediction model based on empirical mode decomposition and investor sentiment, Journal of Computer Applications 43(S1)(2023) 112-118.

[15] Z.-H. Wang, Investor Sentiment Analysis Based on Comment Text for Stock Price Prediction, BCP Business & Management 38(2)(2023) 2710-2716.

[16] X.-Y. Fan, J.-S. Chen, Stock Price Forecasting in Real Estate Industry Based on Investor Sentiment, Frontiers in Business, Economics and Management 6(3)(2022) 54-59.

[17] Y.-W. Yu, S.-S. Wang, L.-J. Zhang, Stock price forecasting based on BP neural network model of network public opinion, in: Proc. 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017.

[18] S.-H. Wu, Y.-L. Liu, Z.-R. Zou, T.-H. Weng, S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis, Connection Science 34(1)(2022) 44-62.