Hindawi
Complexity
Volume 2022, Article ID 7739087, 10 pages
https://doi.org/10.1155/2022/7739087

Research Article

Incorporating Transformers and Attention Networks for Stock Movement Prediction

Yawei Li,[1] **Shuqi Lv**,[2] **Xinghua Liu**,[1] **and Qiuyue Zhang**[1]

[1]*School of Mangement Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China*
[2]*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China*

Correspondence should be addressed to Qiuyue Zhang; 2018309063@stu.sdnu.edu.cn

Received 28 September 2021; Revised 3 February 2022; Accepted 7 February 2022; Published 27 February 2022

Academic Editor: Siew Ann Cheong

Predicting stock movements is a valuable research field that can help investors earn more profits. As with time-series data, the stock market is time-dependent and the value of historical information may decrease over time. Accurate prediction can be achieved by mining valuable information with words on social platforms and further integrating it with actual stock market conditions. However, many methods still cannot effectively dig deep into hidden information, integrate text and stock prices, and ignore the temporal dependence. Therefore, to solve the above problems, we propose a transformer-based attention network framework that uses historical text and stock prices to capture the temporal dependence of financial data. Among them, the transformer model and attention mechanism are used for feature extraction of financial data, which has fewer applications in the financial field, and effective analysis of key information to achieve an accurate prediction. A large number of experiments have proved the effectiveness of our proposed method. The actual simulation experiment verifies that our model has practical application value.

1. Introduction

In recent years, the use of social media information to predict the financial market has attracted the attention of more and more researchers, and some satisfactory experimental results have been achieved. This is due to the fact that social media information contains investor-related attitudes and subjective sentiments towards the financial market, resulting in many investment banks and hedge funds trying to dig out valuable information from media information to help better predict financial markets, which plays a key role in predicting the market. At the same time, the efficient market hypothesis (EMH) introduced by [1] points out that the current price of the asset reflects all the prior information that is immediately available. Therefore, using social media information and the actual price of the current financial market seems to be able to complete the market forecasting more accurately.

The data collected from social media is mainly text, which can be regarded as a natural language processing problem (NLP). Currently, commonly used texts are mainly public news and Twitter, and Twitter is more time-sensitive than public news. In addition, there is a large amount of data reflecting market sentiment in social media, which, of course, or indirectly, affects the trading psychology of investors and thus has an impact on stock prices. Therefore, many recent studies only use tweets to predict stock trends and at the same time explore the characteristic information of these texts at a deeper level [2, 3]. However, due to the time series characteristics of the financial market, historical data at different times has different impacts on the target trading day.

Based on the abovementioned research, we can find that few research studies integrate text, and stock prices to realize stock market forecasts. Experiments show that the method of fusing tweets and actual stock prices with a small time window is superior to the prediction results generated by using text or stock prices alone [4, 5]. This reflects that the wisdom of the crowd comes from the interaction between individuals and the gathering of group opinions. The collective judgment of the group may be much better than the judgment made by a single person. However, currently,

there are few studies using this method to predict the stock market; therefore, it is valuable to propose a research method that integrates social media information and stock prices.

Considering the current results, we still face some challenges: first, the stock market is regarded as time-series data, so the data is temporal dependent, and the value of historical information may decrease over time. In addition, the use of tweets and historical stock prices and other financial data to more accurately analyze and predict the rise and fall of stocks is still a problem worthy of in-depth discussion. In order to effectively solve the above problems, we propose a novel transformer encoder attention (TEA) network architecture, which is a network framework for deep extraction of financial data features and further integration, including a *feature extractor* and a *concatenation processor*, where the historical data uses the information of five calendar days from the target trading day as the training data. In addition, transformer is used to extract text features in-depth, and the attention mechanisms are used to obtain key information from financial data. Through the above framework, one can effectively predict the rise and fall of the target trading day.

For comprehensive experiments, we chose several evaluation metrics to verify the effectiveness of the proposed method in various situations. First, we investigated the feasibility of TEA by comparing its predictive performance with that of other baseline models. Second, we conducted an ablation study of the integrated TEA framework, including ablation of the input data and the model components. Then, visual experiments are presented to demonstrate the effects of various attention mechanisms. Through the performance comparison with related methods, the effectiveness of the method proposed in this paper is fully demonstrated. Furthermore, stock trading simulations are used to illustrate the potential application of the proposed model in actual market trading. Finally, we analyze the performance of the TEA model under special circumstances to detect weaknesses in this method.

In summary, the contributions of our work include the following:

(i) A network framework with a small-time window is proposed, which can effectively solve the problem of temporal dependence of financial data faced in the current stock movement prediction and realize the forecast with the help of historical data closer to the target trading day.

(ii) We introduce the fusion of the transformer model and various attention mechanisms for stock movement prediction. Among them, the transformer model is used to extract deep features of text and uses multiple attention mechanisms to capture dependencies and obtain key information.

(iii) In experiments, the proposed TEA method was found to be significantly superior to several state-of-the-art baselines, demonstrating that the proposed method is more practical than these baselines and can be successfully applied in the financial industry.

The rest of this paper is structured as follows: after this brief introduction, in the Related Work section, a survey of existing work is presented, with a focus on predicting stock movements. Detailed descriptions of our proposed framework are provided in the Methodology section. The Experiments section explains how the data used in the experiments were collected and discusses the results. We conclude the paper and identify future directions of research in the Conclusion and Future Work sections.

2. Related Work

In recent years, deep learning (DL)-based stock prediction methods have been extensively studied and satisfactory results have been achieved [5–7]. Vargas et al. [8] use financial news headlines and prices to estimate the movement of a certain day. The main framework uses a traditional convolutional neural network (CNN) and recurrent neural network (RNN) architectures. Sohangir and Wang [9] propose financial forecasting through DL methods, combined with stock Twitter data, to help investors make decisions. The approach proposes a more innovative analytical approach that provides insights for future solutions. Li et al. [10] proposed a unique approach to text-data fusion that was based on LSTM's model, using four different sentiment dictionaries for a five-year stock market prediction experiment on stock technical indicators and text news data from Hong Kong. However, Lee and Kim [11] proposed a capsule network model based on the transformer encoder (CapTE), which uses the transformer encoder to extract the deep semantic features of social media texts, and then capture the structural relationships of these texts through the capsule network. In addition, attention mechanisms have been applied in many fields since their emergence, and they have now been applied to financial market research problems, especially the temporal attention mechanism, to adapt to the time series of financial data and obtain critical information. For example, [12] developed a hybrid attention network (HAN) to predict stock trends based on the sequence of recent related news items, which can significantly increase annual returns. With the help of tweets and stock prices that is based on incorporative attention mechanisms and multilevel local features, Hu et al. [13] proposed a stock price prediction network (SMPN). Wang et al. [14] have conducted an in-depth analysis of the timeliness and target sensitivity of stock investment reviews, the reliability of investors, public stock reviews, and their application in stock trend forecasting. The paper proposes a new framework based on dynamic experts' attitudes for aggregate stock trend forecasting, in which scores mainly use temporal attention to achieve decision-making. From the above research, it can be found that deep learning has shown great potential in the field of financial market forecasting, and it is expected that better results will be achieved in the future.

In summary, the use of the advanced DL framework to analyze financial data is an effective way to improve stock prediction results. The combined framework that we designed requires in-depth extraction of text and stock price features, and on this basis, can effectively integrate the

extracted features of different types of information so as to completely realize the *Accuracy* of prediction and enhance its application in the actual financial market value. However, the current learning framework still has a lot of room for development in integrating text and numerical data. Therefore, we have designed a complete network framework that can process financial time series data and obtain key information from corresponding tweets. Among them, we adopted a transformer architecture that is completely different from the traditional architecture, mainly considering that it is not troubled by long-term dependence problems, avoiding information dependence problems, and thus greatly reducing training time. In addition, a small-time window, that is, the historical data of 5 days from the target trading day, is used for forecasting so as to obtain more valuable information and avoid focusing on too many resources. Correspondingly, we use an improved temporal attention mechanism to analyze financial data from the perspective of dependence and information content to extract features that are more useful for predicting the stock market. In addition, we adopted an auxiliary forecasting strategy, that is, using previous forecasts to assist in determining the trend of the target trading day. In general, the abovementioned deep learning framework is used to realize the fusion and prediction of text and stock prices so as to be applied to actual stock prediction scenarios.

3. Methodology

In this section, we first summarize the stock movement prediction into specific problems. Then, we put forward our framework in detail. Therefore, we propose the TEA architecture, as shown in Figure 1, which consists of feature extraction and concatenation processing parts.

3.1. Problem Formulation. We wish to predict the movement of stock s on trading day td. We use tweet corpus T and the historical prices in a lag period [d-&d, d-1], where &d is a fixed lag size. The output is a judgment on the binary movement direction, where 1 denotes rise and 0 denotes fall and it is expressed as

$$y = 1\left(p_{td}^{c} > p_{td-1}^{c}\right), \tag{1}$$

where p_{td}^{c} denotes the adjusted closing price, which is adjusted in response to actions that affect the movement of the stock, such as dividends and splits. The adjusted closing price has also been used for stock movement prediction prior to our work [15].

4. Proposed Model: TEA

In order to dig deeper into the valuable information contained in the Twitter text, we use the transformer proposed by Google to obtain text features. Then use the attention mechanism to process the extracted feature information and capture relatively more critical information. At the same time, the normalization strategy is used to deal with the stock price to prevent the stock fluctuation from affecting the stability of the experimental results. The fusion of the two is used as the input to LSTM and the temporal attention mechanism to finally achieve accurate prediction.

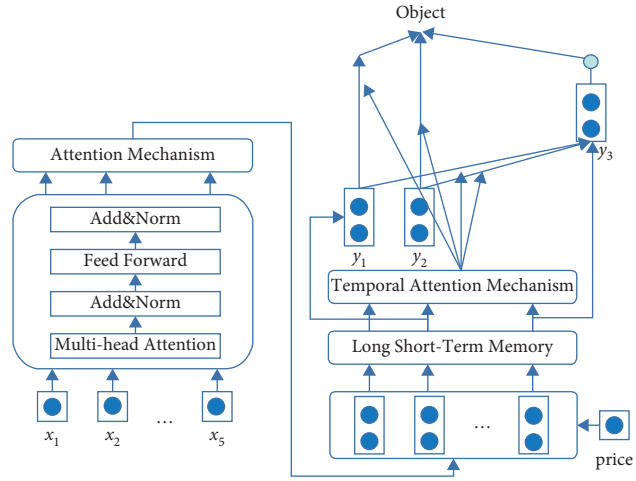

Figure 1: The architecture of TEA.

4.1. Feature Extractor. Here, we will introduce in detail the composition of the feature extraction part, including the processing of text and stock prices. As shown in Figure 2, each layer of the transformer encoder is composed of two submodels: one is a multihead attention mechanism composed of multiple attention mechanisms, the other is a common feedforward neural network, and two addition and normalization (add and norm) are included. The stock price processing process is relatively simple, we mainly adopt the normalization strategy.

4.1.1. Multihead Attention. The essence of multihead is multiple independent attention calculations, which serve as an integrated function to prevent overfitting. Multiple queries are used to calculate multiple pieces of information from the input information in parallel. Each attention focuses on a different part of the input information. In the implementation process, we need to create a query, key, and value vector, and then multiply the word embedding to create three matrices, namely, query (Q), key (K), and value (V). Q, K, and V are first subjected to a linear transformation and then input to the scaling dot product attention. Note that there will be h times. In fact, it is the so-called multihead, and each time it is counted as one head. Moreover, the parameter W for the linear transformation of Q, K, and V is different each time. Then, the h times of scaling dot product attention results are spliced, and the value obtained by performing a linear transformation is used as the result of multihead attention. It can be seen that the difference in the multiheaded attention proposed by Google is that h calculations are performed instead of only one. On the whole, the advantage of the multihead attention mechanism is that it expands the ability of the attention model to pay attention to different location information and provides multiple "representation subspaces" in the attention layer. The matrix

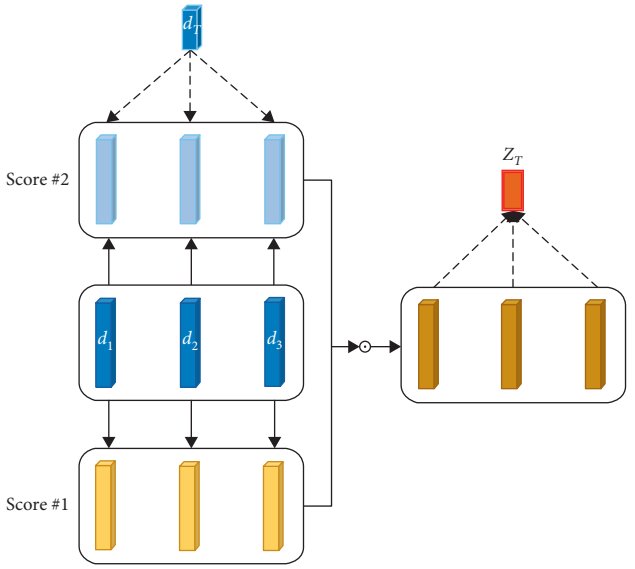

Figure 2: The architecture of temporal attention.

of outputs and the framework are shown in formula (2), and the multihead attention function as follows:

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

The multihead attention function is as follows:

$$\text{Multi Head}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O, \tag{3}$$

where $W^O \in R^{hd_v \times d_{\text{model}}}$ represents the weight matrix. In this work, we employ $h = 5$ parallel attention layers or heads. For each of these, we use $d_k = d_v = d_{\text{model}}/h = 10$.

4.1.2. Addition and Normalization. The multihead attention mechanism passes through the add and norm layers and then passes to the feedforward NN layer. The roles of add and norm are to standardize data and constrain the ambiguity caused by each word in the expression space.

4.1.3. Feedforward NN. Each submodel in the encoder contains a fully connected feedforward neural network (FNN) layer that acts equally on each position. It is made up of two linear transformations with a ReLU activation in between. It is expressed as

$$FFN = \max\left(0, xW^1 + b^1\right)W^2 + b^2. \tag{4}$$

While the linear transformations are the same for all different positions, different parameters are used between different layers. The dimensionality of both the input and the output is $d_{\text{model}} = 50$.

The message embeddings of the tdth trading day are combined to construct an information matrix $M \in R^{d_{\text{model}} \times K}$. The above is the entire feature extraction process, but we also need an attention mechanism to balance the acquired feature information. Specifically, we project the matrix M to u to generate the normalized attention weight, which is achieved with the help of the softmax function.

$$\begin{aligned} u &= \text{soft max}\left(w_u^T \tanh\left(W_m^M\right)\right), \\ c &= Mu^T, \end{aligned} \tag{5}$$

where $W_u \in R^{d_{\text{model}} \times d_{\text{model}}}$ and $W_m \in R^{d_{\text{model}} \times 1}$ are parameters.

In fact, in terms of finance, historical stock prices more truly reflect the conditions of the stock market. Therefore, we have combined the text and real market stock prices. When dealing with stock prices, we need to consider the continuous changes in prices rather than their simple true values. Therefore, instead of directly inputting the original value of td on the trading day into the network framework, we use a normalized strategy to obtain the adjusted closing price. The price adjustment formula is shown as follows:

$$\begin{aligned} P_{td} &= \left[p_{td}^c, p_{td}^h, p_{td}^l\right], \\ P_a &= \frac{p_{td}}{p_{td-1}} - 1, \end{aligned} \tag{6}$$

where p_{td}^c, p_{td}^h, and p_{td}^l denote the closing price, highest price, and lowest price vectors, respectively.

4.2. Concatenation Processor. The abovementioned processing of tweets and stock prices is parallel, and the integration of different types of data is a problem worth exploring. In this article, we use the commonly used fusion method, that is, concatenation, and the data after the fusion of the two are used as the input of LSTM.

$$x = [c, p_a]. \tag{7}$$

LSTM: taking into account the nature of financial time series data, we use LSTM to process the characteristics of the fusion of text and stock prices. LSTM has improved the original RNN and solved the problems of gradient disappearance and long-term dependence. The LSTM unit includes gates and cell states. The cell state is controlled by three gates: the forget gate, the input gate, and the output gate. They cooperate to output the required features. The formulas (8)–(13) describe the implementation process of LSTM and are expressed as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{8}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \tag{9}$$

$$\tilde{C}_t = \tan h\left(W_C \cdot [h_{t-1}, x_t] + b_C\right), \tag{10}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{11}$$

$$O_t = \sigma\left(W_O \cdot [h_{t-1}, x_t] + b_O\right), \tag{12}$$

$$h_t = O_t * \tan h\left(C_t\right). \tag{13}$$

LSTM has been widely used in the field of financial market forecasting and has achieved good experimental results. Therefore, we have adopted this method in the framework proposed in this article. However, the fused features need to adopt a global attention mechanism to analyze the features so as to achieve the prediction task more accurately.

4.2.1. Temporal Attention. Based on the framework we have proposed, it is possible to predict the trend of the target trading day as well as the fluctuations of other trading days during the lag period, which is conducive to the prediction of the target trading day and provides auxiliary information. Therefore, we have adopted a time-series attention mechanism to obtain important information to achieve predictions on targets and other trading days.

Temporal attention is widely used in financial market analysis tasks, and some researchers have improved this mechanism on this basis to adapt to the overall framework. Therefore, we use the information score and dependency score to calculate the contribution weight of the data and further integrate the input that conforms to the characteristics to form the temporal attention model. The structure of temporal attention is shown in Figure 2.

$$d_t = \tan h\left(W_d\left[x, h_t\right] + b_d\right), \tag{14}$$

where W_d denotes the weight matrix and b_d is the bias. The results of the temporal attention model are generated as expressed in

$$v_i' = w_i^T \tanh\left(W_d, i^D\right), \tag{15}$$

$$v_m' = d_T^T \tanh\left(W_{d,m}\right), \tag{16}$$

$$v = \text{softmax}\left(v_i' \odot v_m'\right), \tag{17}$$

$$z_t = \text{softmax}\left(W_y d_t + b_y\right), t < T, \tag{18}$$

$$z_T = \text{softmax}\left(W_T\left[Zv^T, d^T\right] = b^T\right), \tag{19}$$

where W_y and W_T are weight matrices, b_y and b_T are biases, and T represents the set of prediction results for all auxiliary trading days.

5. Experiments

5.1. Dataset and Preprocessing. We validated the performance of our proposed model on three different datasets, including dataset 1, dataset 2, and dataset 3. Dataset 1 used in our work includes tweets and stock prices, which can be found on GitHub. As we all know, stock traders often express their personal opinions on social platforms such as Twitter. Based on the popularity of discussions on the platform, we selected the 88 highest-ranked stocks from January 1, 2014, to January 1, 2016. Dataset 2 is similar to dataset 1, which is composed of tweets and stock prices and includes 47 popular stocks for the time period from January 1, 2017, to January 1, 2018. We also consider dataset 3, which is composed of news headlines and stock prices of 10 stocks for the time period from January 1, 2018, to January 1, 2019.

As we know, different stocks in the stock market have different ups and downs, and some data will only show small changes, which may affect our judgment on the results of the experiment. Therefore, we learn from the method proposed by Hu et al. [12] to specify the upper and lower limits of stock price changes for the stock trend prediction task and set specific thresholds of −0.5% and 0.55%. For samples with motion ratios of ≤ −0.5% and >0.55%, use 0 and 1 to indicate decline and rise, respectively, and delete these data. The remaining dataset is divided by time.

5.2. Training Setup. For the experiments we performed, we set the batch size to 32. The GloVe embedding algorithm is used, and the maximum historical calendar days is set to 5. The traditional Adam optimizer is used when training the model, and the initial learning rate is 0.001. Furthermore, the decay rate is set to 0.96 and the decay step is set to 100.

5.3. Measures of Prediction Performance. We use two indicators to measure prediction performance, namely, *Accuracy* and the Matthews correlation coefficient (MCC). *Accuracy* is widely used in various fields. The MCC is essentially the correlation coefficient between the observed and predicted binary classes. As shown in formula (20), tp, fp, tn, and fn represent the numbers of samples classified as true positives, false positives, true negatives, and false negatives, respectively. The MCC is calculated as follows:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fb)(tn + fn)}}. \tag{20}$$

6. Verifying the Effectiveness of TEA

The experiments are divided into two sets: the purpose of one set of experiments is to verify the feasibility of the proposed model by comparing the performance of TEA with that of baseline models, while the purpose of the other is to analyze all the components of TEA in detail by constructing variants of TEA, namely, an ablation study.

6.1. Comparison with Baseline Models. We have selected several classical and representative baseline methods for comparison with TEA. These models have many similarities, and they all use basic NN architectures to solve the stock market prediction problem:

(i) ARIMA: an advanced technical analysis method that uses only price signals, called the autoregressive integrated moving average method [16].

(ii) TSLDA: a classical model for predicting stock price movements using sentiments on social media, which can capture a theme and the sentiment regarding that theme simultaneously [17].

(iii) HAN: a hybrid attention network for predicting stock trends based on sequences of recent related news items by imitating the learning process of human beings. This model includes news-level attention and temporal attention mechanisms, which are used to focus on key information in the news [12].

(iv) CH-RNN: a novel cross-modal attention-based hybrid recurrent neural network, which consists of two submodules: one makes use of an improved RNN structure to obtain trend representations for different stocks, and the other uses an RNN to model social texts [5].

(v) StockNet: an NN model that uses a VAE to encode input stock data to capture their randomness and analyze the importance of different time steps using temporal attention. The dataset used is the same as that used for our model [4].

(vi) Adv-LSTM: a novel NN prediction model with adversarial training, which is highly expressive for sequential data. This model includes a feature mapping layer, an LSTM layer, a temporal attention mechanism, and a prediction layer; it can extract deep text features and capture the dependencies between texts. Again, the dataset used is the same as that used for our model [7].

(vii) CapTE: a state-of-the-art DL network model that uses a transformer encoder to extract deep semantic features of social media texts and then captures the structural relationships of these texts through a capsule network [2].

(viii) SMPN: a stock movement prediction network (SMPN) with an incorporative attention mechanism has also been proposed. The core innovation is to combine the attention mechanism used for extracting local and contextual features with local semantics that are used to clean up relatively less important contextual embeddings. Thus, the noise in the high-level representation of the structure is effectively reduced, and the performance is improved [13].

(ix) Model1: this paper presents a method to represent price data with a technical index, which uses sentiment vectors to represent news text and sentiment analysis to represent news text. The hierarchical deep learning model is established to learn the sequential information in the market snapshot sequence composed of technical indicators and news sentiment [10].

(x) CNN-BiLSTM-AM: this paper proposes a CNN-BiLSTM-AM method to predict the stock closing price of the next day, which is composed of convolutional neural networks (CNN), bidirectional long-short-term memory (BiLSTM), and an attention mechanism (AM) [18].

(xi) LSTM-RGCN: this paper proposes a more practical objective to predict the overnight stock movement between the previous close price and the open price by making use of the connection among stocks, which models the connection among stocks with their correlation matrix [19].

(xii) FOCUS: this paper proposes a fuzzy 8-momentum inverse uncertain feature system for the 9-classification and quantification of stock features. Features are first identified using random 10 trades, stop-loss, and take-profit mechanisms 11, and then quantified using a novel profitability index 12 with a type 2 fuzzy set module [20].

(xiii) MAFN: it proposes a multihead attention fusion network (MAFN) to exploit aspect-level semantic information in text to improve prediction performance, which adopts an encoder-decoder framework to automatically learn aspect-level text representations through different attention heads [21].

As shown in Table 1, TEA achieved better results than the baseline model in all cases. Compared with the baseline, TEA showed an improvement of 21.13% and 105.63% on dataset 1 in *Accuracy* and *MCC*, respectively. We can see that CapTE has the highest score among all baselines. The histogram in Figure 3 more intuitively shows the changes in *Accuracy* and *MCC*, indicating that the method has achieved certain progress. Since predicting stock trends is a challenging task, even small improvements may bring huge profits. Therefore, for binary stock trend prediction [17], the *Accuracy* of 56% is generally considered to be satisfactory results. In general, the experimental results of the other five baseline models are satisfactory.

Next, we compare the baseline models and TEA on the composition of the framework. First of all, StockNet and Adv-LSTM adopt the bidirectional gated recurrent unit (BGRU) and LSTM structure, respectively, in the text feature extraction stage, while TEA adopts the transformer encoder; the subsequent architectures are basically the same. The results show that the transformer-based method has significant performance advantages over the traditional RNN method. At the same time, CapTE also uses the transformer encoder to extract deep text features. Overall, the experimental results of this method are indeed significantly better than other baseline methods. However, TEA's results are still better than other transformer-based methods, mainly because CapTE's source data only uses social media text information, without considering real stock market prices. In addition, by comparing the overall prediction results of the baseline model, it can be seen that there is still a lot of room for improvement in the performance of the HAN model and the Adv-LSTM model. The fundamental reason is the lack of practical applicability of the submodel and the onesidedness of reflecting emotions. It can be seen that choosing a suitable submodel has a crucial influence on the experimental results. Finally, the experimental results of the CH-RNN model based on RNN are still quite satisfactory. This model uses a variety of different types of attention mechanisms as the mainframe to conduct bidirectional analysis of the target text, thus demonstrating the advantages of the attention mechanism in obtaining key information. SMPN also adopts

Table 1: Performance of baselines and TEA in *Accuracy* and MCC.

	Dataset 1		Dataset 2		Dataset 3	
Models	*Accuracy* (%)	MCC	*Accuracy* (%)	MCC	*Accuracy* (%)	MCC
ARIMA	51.39	−0.0205	51.19	-0.0255	51.37	-0.0205
TLSDA	54.07	0.0653	53.88	0.0614	54.10	0.0661
HAN	57.14	0.0723	57.02	0.0693	56.99	0.0711
CH-RNN	59.15	0.0945	59.15	0.0945	59.00	0.0893
StockNet	58.23	0.0807	57.93	0.0787	58.24	0.0804
Adv-LSTM	57.20	0.1483	57.22	0.1395	57.31	0.1501
CapTE	64.22	0.3481	64.18	0.3405	**64.29**	0.3504
SMPN	59.74	0.1298	58.23	0.2542	58.57	0.2133
Model1	58.40	–	58.11	0.2172	56.83	0.1952
CNN-BiLSTM-AM	53.95	0.1351	52.65	0.2130	53.62	0.1757
LSTM-RGCN	58.90	0.2931	59.32	0.2944	59.17	0.2842
FOCUS	57.42	0.1701	57.33	0.1832	57.21	0.2201
TEA	**64.57**	**0.3512**	**65.22**	**0.3411**	64.07	**0.3521**

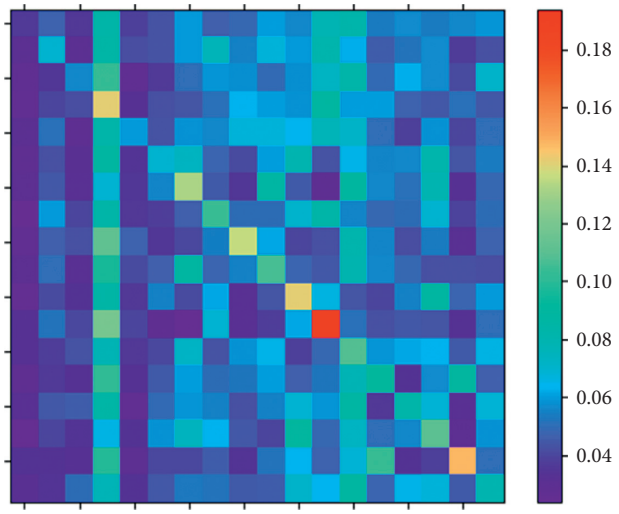

Figure 3: Visualization of attention scores.

the idea of fusing text and prices with the attention mechanism as the main body and adopts the method of concatenation. Similarly, model1 considers news text and integrates emotional information, but the model it uses is a more traditional method and does not deeply integrate the two types of information.

In recent years, there have been some new methods to solve the problem of stock prediction, especially when combined with knowledge graphs, we can see that LSTM-RGCN obtained great results using the graph. And the model of CNN-BiLSTM-AM combined the traditional method and deep learning model to predict the stock. Maybe how to combine them better is a problem to explore. And also, we find that FOCUS used the fuzzy system to achieve trading, it seems that this kind of method is a novel direction. However, the results of the above methods behave poorer than our method, there are some reasons: first, there are differences in extracting features, which may play an important influence on prediction; second, historical data and time window are also key factors. It is observed that our method obtained satisfactory results on three datasets, which demonstrate the great effect.

6.2. Ablation Study. To perform a detailed analysis of all of the primary components of TEA, in addition to the fully equipped TEA, we also constructed the following four variants. Among them, "TEA (W/O tweets)" means that the input data does not include the tweet corpus. Moreover, we propose variants with alternative parameter values in which the length of the lag period is modified to 7 and 10 days, named TEA (lag_size_7) and TEA (lag_size_10), respectively:

(i) TEA (W/O tweets): TEA with the original parameters using only historical prices as input data

(ii) TEA (W/O price): TEA with the original parameters using only tweets as input data

(iii) TEA (lag_size_7): TEA with alternative parameter settings corresponding to a lag period of 7 calendar days

(iv) TEA (lag_size_10): TEA with alternative parameter settings corresponding to a lag period of 10 calendar days

Table 2 shows the performance of the different TEA variants. We can see that TEA (W/O Tweet) produces the worst results, indicating that the text data from social media does include some valuable information. Even compared with the aforementioned baseline model, TEA (W/O Tweet) is not satisfactory, which shows the necessity of extracting text features as input. By contrast, TEA (W/O price) produces exceptionally competitive results, similar to those of TEA. Consequently, it can be concluded that social media texts contain a large amount of market information, and several studies have shown that this kind of information is useful when predicting stock prices and even for other financial tasks [22]. The performance results of TEA (W/O tweets) and TEA (W/O price) confirm the positive effects of tweets and historical prices, respectively, for stock movement prediction.

In TEA, the trading day is the basic forecast unit, so the choice of time window has a more direct impact on the experimental results. However, when we use three calendar days to forecast, there are restrictions on the existence of multiple trading days in the lag period, such as monday's

Table 2: Performance of TEA variations in *Accuracy* and MCC.

Models	Dataset1		Dataset 2		Dataset 3	
	Accuracy (%)	MCC	*Accuracy* (%)	MCC	*Accuracy* (%)	MCC
TEA (W/O tweets)	58.19	0.1597	57.12	0.1429	58.03	0.1631
TEA (W/O price)	63.67	0.3212	63.17	0.3008	63.82	0.3119
TEA (lag_size_7)	63.99	0.3379	63.21	0.2899	63.15	0.2822
TEA (lag_size_10)	61.97	0.1982	60.77	0.1859	60.66	0.1834

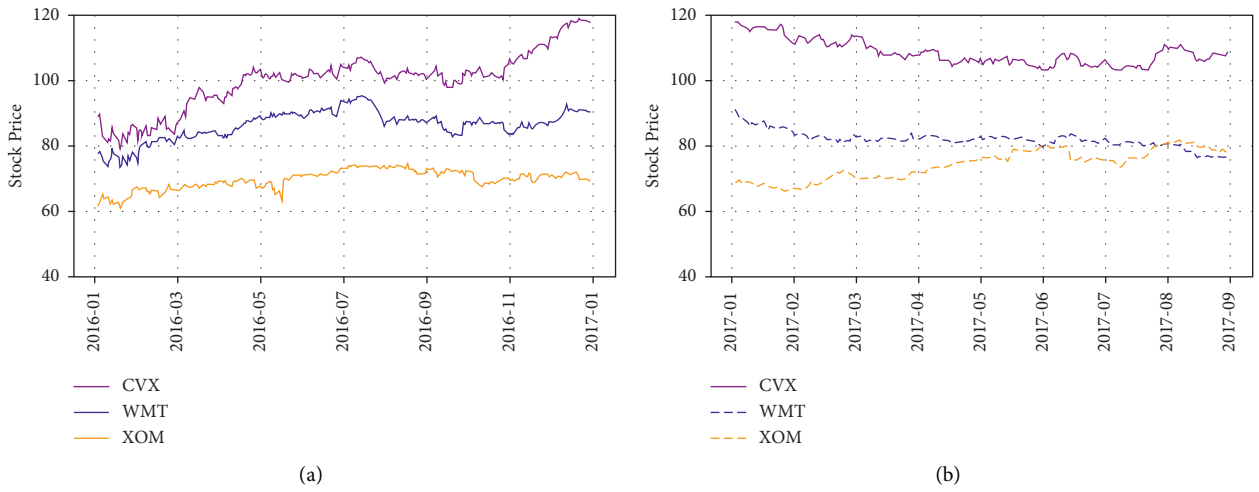

Figure 4: Demonstration of temporal attention.

sports forecast. As shown in Table 2, experiments with a time window of 7 and 10 calendar days did not produce better results than experiments with a lag of 5. There are two main reasons: many random factors affect the stock market, and changes in the target trading day are more likely to be affected by the price of the adjacent trading day. In addition, investor comments often reflect current market conditions and are not suitable for predicting long-term trading behavior.

7. Effects of Attention Mechanisms

To qualitatively analyze the attention mechanisms in our model, we chose to use tweets and stock prices to represent the capabilities of the multihead attention and temporal attention mechanisms, respectively.

7.1. Visualization of Multihead Attention. Since the text is part of the input vector, we explore the attention weight of each text when it is input, and observe the role of the corresponding module in the framework in a visual form. The visual effect of the attention mechanism could intuitively show its effect, which demonstrates the effect of the whole model. Therefore, it is crucial to explore the attention mechanism in the deep learning model. Especially for our model, we used several different attention mechanisms. In Figure 3, we selected one sentence of opinions about one stock and then used the multihead attention mechanism to analyze it. It shows the attention distribution of a randomly selected tweet, where the darker the color, the more important the corresponding word is. As can be seen from Figure 3, data with lower attention scores obviously cannot provide valuable information, while higher data are clearly used to predict market trends. Therefore, the text processing method we adopted is feasible and can significantly improve performance.

7.2. Demonstration of Temporal Attention. In order to further illustrate the impact of the information used in the improved temporal attention mechanism and the dependent scores on the actual stock market, we observe the relationship between stock price changes and tweet information. In Figure 4, we have plotted the rise and fall of three stocks over time. It can be concluded that the trend of stocks is not only affected by historical stock prices but also by tweets, which directly proves that the improved timing attention mechanism we proposed is comprehensive. In fact, investors can carefully synthesize relevant investor comments and prices to better assess their impact on the underlying stocks. Therefore, the ideal framework for simulating this analysis process should integrate and interpret such information over a continuous period of time, rather than analyzing these factors separately.

8. Error Analysis

In this article, we compare the performance of TEA and CapTE and analyze the situations where TEA produces wrong results but CapTE is correct. There are two scenarios: firstly, if a trader writes a tweet with negative

sentiment, for example, the tweet "Mobile phones are depreciating faster than Vanke A is falling," the comment was actually talking about the devaluation of mobile phones, but the trader's sentiment extended to the stock, leading to a malicious comment about the corresponding stock. Secondly, comments about market issues from long ago may appear on Weibo. For example, the financial tsunami discussed in "The 2008 US-induced financial tsunami caused Chinese stock prices to fall by almost 80%" occurred in 2008 and is not relevant to the current stock market, yet in this case, it would be difficult for our model to obtain correct predictions without introducing relevant information.

9. Conclusion and Future Work

In this paper, we have proposed a novel DL model called TEA, which can use the historical stock prices from five calendar days in combination with social media tweet representations to predict stock movements by means of a transformer encoder and attention mechanisms. To solve the problems of temporal dependence in financial data and insufficient effectiveness in fusing information from tweets and stock prices, an architecture consisting of a *feature extractor* and a *concatenation processor* is adopted. For the *feature extractor*, the structure consists of a transformer encoder, attention mechanisms, and a normalization strategy to effectively extract features and learn key information through relevant reviews of tweets and preprocessing of stock prices. The *concatenation processor* then processes the fused features to capture the temporal dependence. The overall framework realizes effective processing and analysis of tweets and stock prices to improve the *Accuracy* of stock movement prediction.

In this study, five main sets of experiments were performed to verify the validity of the proposed model. First, we compared the prediction performance with that of various baseline models to investigate model feasibility. Second, four different variants of the TEA model were constructed and tested to analyze the impact of different components. Third, the effects of the attention mechanisms were intuitively investigated to confirm the contributions of the multihead attention and temporal attention mechanisms to the overall performance. Then, trading simulations were performed as a profitability test. Finally, we performed an error analysis to analyze TEA from various perspectives. The results indicate that the proposed model successfully achieves the stock movement prediction task with satisfactory experimental performance. Through comparisons with state-of-the-art methods, we find that the proposed method is more suitable for practical application and can help traders avoid financial risks and make more favorable decisions.

Data Availability

Data used in the study are available at the following link: https://github.com/yumoxu/stocknet-dataset.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61972227), the Natural Science Foundation of Shandong Province (Grant no. ZR2019MF051 and ZR201808160102), the Focus on Research and Development Plan in Shandong Province (Grant no. 2020CXGC010110), and the Independent Training and Innovation Team of Shandong Province (Grant no. 2020GXRC016) and in part by the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

References

[1] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll, "The adjustment of stock prices to new information," *International Economic Review*, vol. 10, pp. 34–105, 1969.

[2] J. Liu, X. Liu, H. Lin et al., "Transformer-based capsule network for stock movements prediction," in *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pp. 66–73, Macao, China, August 2019.

[3] D. Araci, "FinBERT," *Financial Sentiment Analysis with Pre-trained Language Models*, 2019, https://arxiv.org/abs/1908.10063v1.

[4] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pp. 1970–1979, Melbourne, Australia, July 2018.

[5] H. Wu, W. Zhang, W. Shen, and J. Wang, "Hybrid deep sequential modeling for social text-driven stock prediction," in *Proceedings of the 2018 Association for Computing Machinery*, pp. 1627–1630, Torino, Italy, October 2018.

[6] Q. Ding, S. Wu, H. Sun, J. Guo, and J. Guo, "Hierarchical multi-scale Gaussian transformer for stock movement prediction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pp. 4640–4646, Christian, Bessiere, July 2020.

[7] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T. S. Chua, "Temporal relational ranking for stock prediction," *ACM T INFORM SYST*, vol. 37, 2019.

[8] M. R. Vargas, C. E. M. Dos Anjos, G. L. G. Bichara, and A. G. Evsukoff, "Deep leaming for stock market prediction using technical indicators and financial news articles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 971–978, New Orleans, Louisiana, USA, July 2018.

[9] S. Sohangir and D. Wang, "Finding expert authors in financial forum using deep learning methods," in *Proceedings of the 2018 Second IEEE International Conference on Robotic Computing (IRC)*, pp. 399–402, Laguna Hills, CA, USA, February 2018.

[10] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong," *Information Processing & Management*, vol. 57, no. 5, Article ID 102212, 2020.

[11] S. W. Lee and H. Y. Kim, "Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation," *Expert Systems with Applications*, vol. 161, Article ID 113704, 2019.

[12] Z. Hu, W. Liu, J. Bian, X. Liu, and T. Y. Liu, "Listening to chaotic whispers: a deep learning framework for news-oriented Stock trend prediction," in *Proceedings of the WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 261–269, Los Angeles, CA, USA, February 2018.

[13] H. Xu, L. Chai, Z. Luo, and S. Li, "Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices," *Neurocomputing*, vol. 418, pp. 326–339, 2020.

[14] H. Wang, T. Wang, and Y. Li, "Incorporating expert-based investment opinion signals in stock prediction: a deep learning framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 971–978, New York, USA, February 2020.

[15] B. Xie, R. J. Passonneau, L. Wu, and G. G. Creamer, "Semantic frames to predict stock price movement," in *Proceedings of the ACL 2013 -51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 873–883, Sofia, Bulgaria, August 2013.

[16] L. S Goodman, "Book Reviews 2," pp. 314-315, 2016, http://www.jstor.org/stable/3150192.

[17] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 873–883, Beijing, China, July 2015.

[18] W. Lu, J. Li, J. Wang, and L Qin, "A CNN-BiLSTM-AM method for stock price prediction," *Neural Computing and Applications*, vol. 33, pp. 1–13, 2020.

[19] W. Li, R. Bao, K. Harimoto, D Chen, J Xu, and Q Su, "Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20*, Yokohama, Japan, January 2020.

[20] M. E. Wu, J. H. Syu, J. Lin, and J. M Ho, "Effective fuzzy system for qualifying the characteristics of stocks by random trading," *IEEE Transactions on Fuzzy Systems*, 2021.

[21] N. Tang, Y. Shen, and J. Yao, "Learning to fuse multiple semantic aspects from rich texts for stock price prediction," in *Proceedings of the International Conference on Web Information Systems Engineering*, pp. 65–81, Springer, Hong Kong, China, November 2019.

[22] O. Bustos and A. Pomares-Quimbaya, "Stock market movement forecast: a Systematic review," *Expert Systems with Applications*, vol. 156, Article ID 113464, 2020.