IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 8, 2018, accepted December 3, 2018, date of publication December 12, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886367

A Numerical-Based Attention Method for Stock Market Prediction With Dual Information

GUANG LIU AND XIAOJIE WANG

Center for Intelligence of Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Xiaojie Wang (xjwang@bupt.edu.cn)

This paper is supported by NSFC (No. 61273365), NSSFC (2016ZDA055), 111 Project (No. B08004), Beijing Advanced Innovation Center for Imaging Technology, Engineering Research Center of Information Networks of MOE, China. The authors would like to thank the anonymous reviewers for their valuable comments.

ABSTRACT Stock market prediction is of great importance for financial analysis. Traditionally, many studies only use the news or numerical data for the stock market prediction. In the recent years, in order to explore their complementary, some studies have been conducted to equally treat dual sources of information. However, numerical data often play a much more important role compared with the news. In addition, the existing simple combination cannot exploit their complementarity. In this paper, we propose a numerical-based attention (NBA) method for dual sources stock market prediction. Our major contributions are summarized as follows. First, we propose an attention-based method to effectively exploit the complementarity between news and numerical data in predicting the stock prices. The stock trend information hidden in the news is transformed into the importance distribution of numerical data. Consequently, the news is encoded to guide the selection of numerical data. Our method can effectively filter the noise and make full use of the trend information in news. Then, in order to evaluate our NBA model, we collect news corpus and numerical data to build three datasets from two sources: the China Security Index 300 (CSI300) and the Standard & Poor's 500 (S&P500). Extensive experiments are conducted, showing that our NBA is superior to previous models in dual sources stock price prediction.

INDEX TERMS Deep learning, machine learning, natural language processing, prediction methods, stock markets.

I. INTRODUCTION

Stock market prediction aims to determine the future value of a company stock traded on an exchange. Reliable prediction of future stock prices can yield significant profits. Many researchers have adapted the news and numerical data for stock market prediction [3]–[5], [8], [13]–[15], [18], [33], [36], [44].

According to the number of information sources, the stock market prediction methods can be grouped into two categories: single-source methods and dual-source ones.

In the single-source methods, Qin *et al.* [36] use only numerical data to predict the minute price of NASDAQ100 index. Chong *et al.* [8] use only numerical data to explore the effects of deep learning method in stock market prediction. Xie *et al.* [44] use the semantic information extracted from the news for predicting the direction of stock movement. Ding *et al.* [11] extract event information from news titles for predicting the trend of related stocks. All these studies only use information from a single source.

The Efficient Market Hypothesis (EMH) [30] indicates that all relevant information will be reflected in the stock price. Information from different sources can complement each other and affect the stock price. Therefore, a few studies start to use both news and numerical data to predict the stock market [10], [24], [41], [46].

Dual-source methods focus on effective representations for news and on capturing the temporal correlation within information. Schumaker and Chen [38] have studied the effectiveness of different textual representations of news, Bag of Words, Noun Phrases, and Named Entities combined with current stock price to predict stock price 20 minutes after the news report is publicly available. They claim that the article terms combined with the stock price at the time of news release have the best performance in closeness and profit. In their research, the sparse representations of textual information, which suffer from the curse of dimension, are used. In order to solve this sparsity in news representation, Akita *et al.* [1] explore the use of Paragraph Vector (PV) and

numerical data in predicting the price of companies listed on the Tokyo Stock Exchange, which can make a profit from their experiment settings. However, they use the Long Short-Term Memory (LSTM) as the predictor, which cannot manage the long temporal dependency in the target prices. Therefore, Xu and Cohen [45] propose a complex generative model to learn the temporal dependency in future prices and to predict the trend of stock price movement with dual-information.

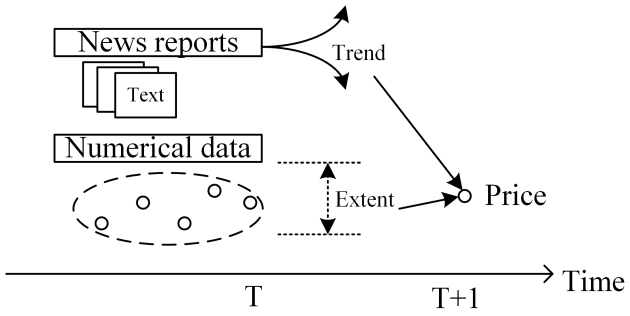


FIGURE 1. Demostration the effects of dual source information.

However, existing methods essentially assume features from different sources are equally important. Thus, their synthesis of information is relatively simple and the complementary between them are ignored [1], [38]. As shown in Figure 1, the ability in exhibiting the extent of stock price movement makes the numerical data play a major role in stock price prediction; and the ability to indicate future trends of related stocks makes news a helpful complement to numerical data. Direct concatenation treats information from news and numerical data equally [27]. And the stock trend in the news and the extent of the stock price in numerical data are not properly organized and combined [23], thereby limiting the model's ability to learn their connections to stock market volatility.

To address the problem in combining dual-information, we propose a Numerical-based Attention (NBA) method for stock price prediction. In our method, the encoded news is used to guide the attention method and select relevant numerical data across all time steps. Hence, the tendency of related stocks from the news is transformed into the importance of numerical data. Benefited from the transforming, information from the news is able to supplement the numerical data with the trend information. The attention mechanism [2] provides an effective means of combining dual-information which can effectively filter the noise and make full use of the trend information in news.

To evaluate our model, we collect news and numerical data to build three datasets from two stock markets: the China stock market and the US stock market. The experimental results show that our NBA is superior to baselines in predicting the stock price with dual-information. We also demonstrate that our NBA can effectively capture the complementary between news and numerical data.

In summary, the main contributions of this work are as follows: 1) We propose an attention-based method to effectively ultilize the complementarity between news and numerical data to predict stock prices. News, which contains the stock trend information, is converted into the importance distribution of the numerical data. Thus, the news gives guidance on the selection of numerical data. The proposed method can effectively reduce the noise and make full use of the trend information in news. 2) Three datasets with both news corpus and numerical data source from the China Security Index 300 (CSI300) and Standard & Poor's 500 (S&P500) are built to evaluate our method. Extensive experimental results show that our NBA outperforms previous models in dual sources stock price prediction.

The remainder of this paper is organized as follow. In Section 2, we illustrate the stock price prediction and sequence-to-sequence method. Then, we describe the architecture of our model in Section 3. In Section 4, the experimental settings are summarized. Next, Section 5 presents the experimental results and analysis of the S&P500 and CSI300 financial news datasets. In the last section, the conclusions are given.

II. RELATED WORKS

A. STOCK TREND PREDICTION

Accurate and reliable prediction of the stock price is of great importance to both investors and researchers. Numerous researches have been devoted to investigating effective methods towards predicting stock price [11], [36]. In terms of the target number, the prediction methods can be classified into short-term methods and long-term methods [39]. Short-term methods, or one-step ahead methods, attempt to forecast price movement one-step ahead [34]. Although short-term methods have yielded encouraging results, there are unable to analyze the impact of a piece of information over a period, such as breaking news. Therefore, long-term methods, or multi-step ahead methods, are proposed to address this problem. The major difference between these two methods is the number of targets [12]. In contrast to short-term methods, long-term methods aim to predict multi-step ahead stock prices. It draws more challenges due to the growing uncertainties arising from various factors, for instance, accumulation of errors and the ignorance of information [39]. In this work, we focus on the long-term prediction of stock prices.

B. TEXTUAL-BASED STOCK PREDICTION

Textual-based stock prediction methods can be divided into the news-oriented methods and social-oriented methods. The social-oriented methods mainly use textual information from social networks, such as Tweet or Weibo. They mainly focus on applying sentiment analysis techniques to social media data to get public moods. Some methods use the models based on the topic model [37]; other methods use lexicon based approaches [6], [22]. However, comparing to news, the acquisition of public sentiment is more

difficult because it requires sufficient coverage to ensure the correct sampling of public sentiment. News is more accessible than social network data. The major source of news-oriented methods is the texts of news. The researches mainly focus on either news sentiment analysis [25], [28], [42] or extracting effective features [26]. However, these methods require certain supervision, which consumes many resources. In recent years, unsupervised feature extraction based on deep learning achieves remarkable results in the field of Nature Language Processing (NLP) [16], [35]. Moreover, unsupervised sentence representation based on deep learning has found many applications in text classification and sentiment classification [20], [32]. As a result, this paper analyses the effects of various unsupervised representations of news based on NLP for stock prediction.

C. SEQ2SEQ METHOD WITH ATTENTION

Sequence-to-sequence method is proposed by Cho *et al.* [7] 2014 and Sutskever *et al.* [40] 2014. The framework often has two components for its architecture. The encoder encodes a variable-length sequence into a fixed-length vector. The decoder decodes the given fixed-length vector back into a variable-length sequence. The model provides a general method to learn the conditional distribution of two sequences, e.g. $p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T)$, where the lengths of input and output sequences T' and T may differ. The hidden state of the encoder at time t is computed by

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, x_t). \tag{1}$$

Here $f(\cdot)$ is a non-linear activation function, x_t is the input at time step t. The hidden state of the decoder at time t is computed by

$$\mathbf{s}_t = f(\mathbf{s}_{t-1}, y_t, \mathbf{c}). \tag{2}$$

Here $\mathbf{c} = \{(c_1, c_2, \cdots\}$ is the summary of the input sequence. The conditional distribution of the next symbol is

$$P(y_t|y_{t-1}, \cdots, y_1, \mathbf{c}) = g(s_t, y_{t-1}, \mathbf{c}) \tag{3}$$

for the given activation functions $f(\cdot)$ and $g(\cdot)$. The model is jointly trained to maximize the conditional log-likelihood

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} p_\theta(y_n|x_n), \tag{4}$$

where θ is the set of the model parameters and each (x_n, y_n) is an (input sequence, output sequence) pair from the training set.

Attention-based method is proposed to manage the long-term dependency in decoding process [2]. In sequence-to-sequence model, the attention mechanism is meant to assign different weights to each part of the input sequence [29]. The context vector $\mathbf{c}_i$ depends on a sequence of annotations $(\mathbf{h}_1, \cdots, \mathbf{h}_T)$ to which an encoder maps the input sequence. Each annotation $\mathbf{h}_i$ contains information about the whole input sequence with a strong focus on the parts surrounding the ith observation of the input sequence. The context vector $\mathbf{c}_i$ is then computed as a weighted sum of these annotations $\mathbf{h}_j$:

$$\mathbf{c}_i = \sum_{j=1}^{T} \alpha_{ij} \mathbf{h}_j. \tag{5}$$

The weight α_{ij} of each annotation $\mathbf{h}_j$ is computed by

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{T} \exp(e_{ij})}, \tag{6}$$

where $e_{ij} = a(\mathbf{s}_{i-1}, \mathbf{h}_j)$ scores how well the input around position j and the output at position i match. The score is based on the decoder hidden state $\mathbf{s}_{i-1}$ and the jth annotation $\mathbf{h}_j$ of the input sentence. In this work, we make some improvements on the original attention-based sequence-to-sequence method to learn the correlation between dual-information and stock market volarity.

III. NUMERICAL-BASED ATTENTION

In this section, we first introduce the notation we use in this work. Then, we present the architecture of the proposed model.

A. NOTATION AND STATEMENT OF PROBLEM

For the given time t, let $\mathbf{x}_{a,i} \in \mathbb{R}^V$ be the numerical vector at time $t - K + i$ and $\mathbf{x}_{b,j} \in \mathbb{R}^L$ be the jth textual information released at time t, where V is the number of variables in numerical information, K is the length of windows size and L is the size of vocabulary. For simplicity, we employ $\mathbf{X}_{a,K} = (\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \cdots, \mathbf{x}_{a,K}) \in \mathbb{R}^{K \times V}$ to denote the input numerical sequence, employ $\mathbf{X}_{b,d} \in \mathbb{R}^{d \times L}$ to denote the set of textual information released at time t and employ $Y_{t,M} = (y_1, y_2, \cdots, y_M) \in \mathbb{R}^M$ to denote the target sequence of stock price from $t + 1$ to $t + M$, where d is the number of textual information, y_i is the ith stock price in the target sequece.

Typically, our model aims to learn a nonlinear mapping to the target sequence $Y_{t,M}$ with

$$Y_{t,M} = F(X_{a,K}, X_{b,d}), \tag{7}$$

where $F(\cdot)$ is a nonlinear mapping function.

B. MODEL ARCHITECTURE

We propose a model with Numerical-based Attention (NBA) for stock market prediction. The architecture of our model is based on a typical sequence-to-sequence model. The dual information encoder converts numerical data and news into the vector representations, and the numerical-based decoder predicts stock prices basing on the representations. In order to learn the complementary between news and numerical information, we introduce a numerical-based attention mechanism in the decoder. It is worth mentioning that we use the feature vector of news to guide the selection of related numerical data across all time steps. In this way, our model is able to select trend-related data to construct the context vector. Then, the context vector accompany with decoder state are fed to a linear layer to make price predictions.

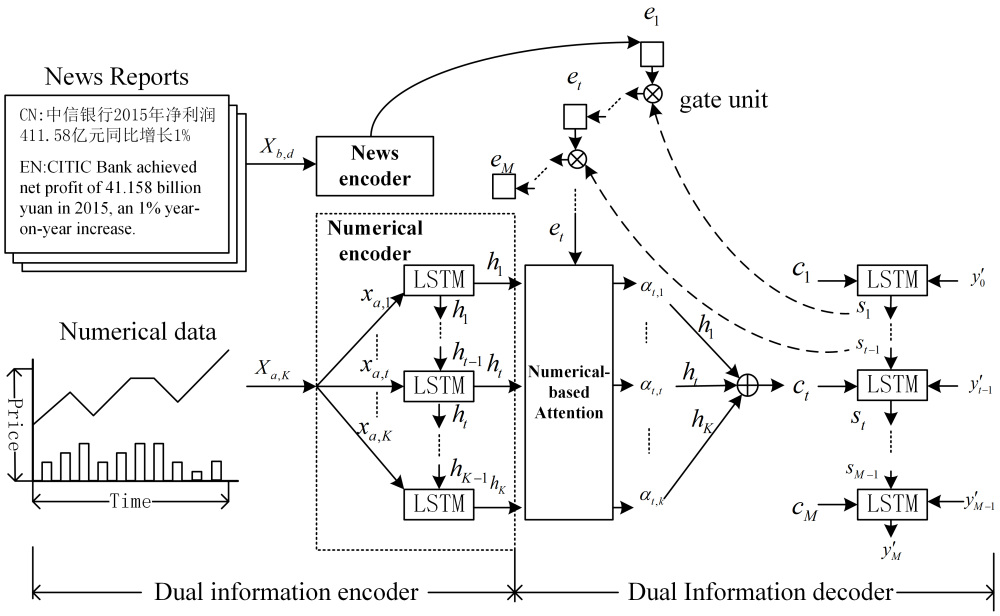


FIGURE 2. The architechture of NBA.

Fig. 2 illustrates the architecture of the model. Our model consists of two major components. The first component is a dual-information encoder, which encodes the news and numerical data and outputs the news embedding and numerical vector. The second component is the numerical-based decoder, which decodes the numerical vector into a sequence of stock prices with numerical-based attention.

1) DUAL INFORMATION ENCODER

The dual-information encoder processes the numerical data and news accordingly. For the numerical data, we use a non-linear function to encode them into a corresponding numerical vector. For news, we use dense representations to construct the news embedding.

a: NUMERICAL ENCODER

In this case, given the numerical data $\mathbf{X}_{a,K} = (\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \cdots, \mathbf{x}_{a,K})$ with $\mathbf{x}_{a,t} \in \mathbb{R}^V$, where K is the length of input sequence. The hidden state of each time-step can be calculated as

$$\mathbf{h}_t = f_{en}(\mathbf{x}_{a,t}, \mathbf{h}_{t-1}; \theta_{en}), \tag{8}$$

where $\mathbf{h}_t \in \mathbb{R}^m$ is the hidden state of the encoder at time t, m is the size of hidden state, $f_{en}(\cdot)$ is a non-linear function, and θ_{en} is the parameters of encoder function. The Recurrent Neural Networks (RNN) is often used as an encoder in Machine Learning Translation [7]. There are many choices in encoding the data sequence. However, the LSTM is one of a few models which can capture the dynamic structure within the sequence. In addition, the LSTM (Hochreiter and Schmidhuber [17]) overcomes the problem of long dependency when encoding the various lengths of sequence into fixed-length vector. As a result, we use the LSTM as our numerical data encode.

Each LSTM unit has a memory cell controlled by three sigmoid gates units. The hidden state $\mathbf{h}_t$ is updated as follows:

$$f_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{9}$$
$$i_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{10}$$
$$o_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{11}$$
$$\mathbf{Q}_t = f_t \odot \mathbf{Q}_{t-1} + i_t \odot \tanh(\mathbf{W}_C[\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{12}$$
$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{Q}_t) \tag{13}$$

where $f_t, i_t, o_t \in \mathbb{R}$ are the sigmoid gate units at time t, $\mathbf{Q}_t \in \mathbb{R}^m$ is the cell state at time t, and $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_o \in \mathbb{R}^{m\times(m+V)}$ are the parameters. In this paper, the bias term are ignored for simplicity. Here, we use θ_{en} to denote the parameters in encoder LSTM.

b: NEWS ENCODER

We only use the headlines of news, which are published in the same time span, to construct news embedding based on unsupervised dense representations, since employing barely news headlines can achieve better performance than using both the headline and content of news in stock movement prediction [11]. Although adapting events extracted from news headlines can achieve higher trend performance, this approach heavily depends on Open Information Extraction (OpenIE) tools, which has very few alternatives in non-English languages. Therefore, we use Paragraph Vector [20], Sequential Denoise AutoEncoder (SDAE) [16], Word2vec [32] and Glove [35] to construct news embedding. For some given news headlines $\mathbf{X}_{b,d} \in \mathbb{R}^{d\times L}$, two different methods can be applied to construct news embedding [1], [11].

1) *Bottom-up method.* As shown in Fig. 3, the bottom-up method averages each word embedding (Word2vec

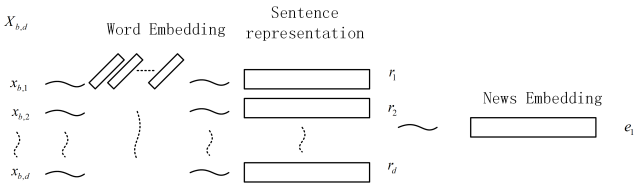


FIGURE 3. The bottom-up method to construct sentence embedding.

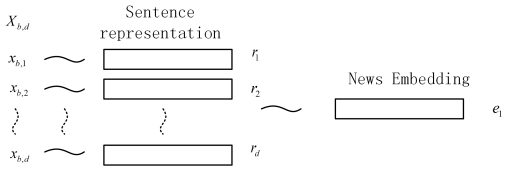


FIGURE 4. The direct method to construct sentence embedding.

or Glove) in a news title as its sentence representation. Then, the averaged sentence representation of all news headlines gives the news embedding $\mathbf{e}_1 \in \mathbb{R}^n$.

2) *Direct method.* As shown in Fig. 4, we first get sentence representation for each news titles. Then, we calculate the news embedding $e_1 \in \mathbb{R}^n$ by averaging the overall sentence representations in the same time span.

2) DUAL INFORMATION DECODER

To predict future stock prices, we use a different LSTM to decode the numerical vector. In order to learn the complementary between news and numerical data, we adapt the temporal attention method in the decoder, which is proposed to address the rapid deterioration of performance as the length of the input sequence increases [36]. The proposed attention-based method transforms information in news into the importance of numerical data. We call this novel attention method the Numerical-Based Attention (NBA). It can adaptively select the trend-related encoder hidden states across all time steps. Specifically, the attention weight of each encoder hidden state at time t is calculated basing on the encoder hidden state $\mathbf{h}_i \in \mathbb{R}^m$ and the news embedding $\mathbf{e}_t \in \mathbb{R}^n$ with

$$\beta_{i,t} = \mathbf{v}_a^T \tanh(\mathbf{W}_a[\mathbf{h}_i; \mathbf{e}_t]), 1 \le i \le K, \tag{14}$$

$$\alpha_{i,t} = \frac{\exp(\beta_{i,t})}{\sum_{j=1}^{K} \exp(\beta_{i,j})}, \tag{15}$$

where $[\mathbf{h}_i; \mathbf{e}_t] \in \mathbb{R}^{m+n}$ is a concatenation of the encoder hidden state and news embedding, $\mathbf{v}_a \in \mathbb{R}^m$ and $\mathbf{W}_a \in \mathbf{R}^{m\times(m+n)}$ are parameters to learn. The vector $\mathbf{e}_t \in \mathbb{R}^n$ is the news embedding calculated by

$$\mathbf{e}_t = \begin{cases} \mathbf{e}_1 & t = 1 \\ \mathbf{e}_{t-1} \odot \sigma(\mathbf{W}_g \mathbf{s}_{t-1}) & t > 1. \end{cases} \tag{16}$$

where $\mathbf{s}_{t-1} \in \mathbb{R}^m$ is the decoder hidden state at time step $t-1$, $\mathbf{W}_g \in \mathbb{R}^{n\times m}$ is the parameters and $\sigma(\cdot)$ is the sigmoid function. The news embedding is gated by the decoder state to select the information passing to the next time step in decode progress. This operation simulates the dynamical effects of news. The hidden state $\mathbf{s}_t$ is updated as:

$$\mathbf{s}_t = f_{de}(\gamma_t, \mathbf{s}_{t-1}; \theta_{de}), \tag{17}$$

where $f_{de}(\cdot)$ is an LSTM unit, θ_{de} denotes the parameters in decoder LSTM, γ_t is the decoder input at time t which is calculated by

$$\gamma_t = \mathbf{W}_\gamma^T[y'_t; \mathbf{c}_t], \tag{18}$$

where $\mathbf{W}_\gamma \in \mathbb{R}^{m+1}$ is the parameter matrix, $\mathbf{c}_t$ denotes the context vector. For a given encoder hidden state $(\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_K)$, the vector $\mathbf{c}_t$ is calculated as follows

$$\mathbf{c}_t = \sum_{i=1}^{K} \alpha_{i,t} \mathbf{h}_i. \tag{19}$$

Note that the attention weight $\alpha_{i,t}$ is guided by news, which exhibits tendency information. Thus, the context vector $\mathbf{c}_t$ is the trend-related weighted average of hidden states across all time steps. The state of decoder is initialized by the last hidden state in encoder. For the first decode step, we use the close price at news release time y'_0 as the input. Other steps in decoding use the predicted value of the previous step as the input. In the upgrade process of decoding, the hidden state is identical to those in (9)-(13). The attention weight $\alpha_{i,t}$ represents the importance of the ith encoder hidden state for the prediction.

In summary, we aim to predict the future price sequence $\mathbf{Y}_{t,M} = (y_1, y_2, \cdots, y_M) \in \mathbb{R}^M$. For the prediction of multi-step ahead prices of a given stock after the related news is released, we use the NBA to approximate the mapping function F so as to obtain the hidden state of the decoder $\mathbf{s}_t$ with the observation of all inputs as well as the previous outputs. Specifically, the predicted price y'_t at time t is calculated by

$$y'_t = F(\mathbf{X}_{a,K}, \mathbf{X}_{b,d}) = \mathbf{v}^T(\mathbf{W}_p^T[\mathbf{s}_t; \mathbf{c}_t]), \tag{20}$$

where $[\mathbf{s}_t; \mathbf{c}_t] \in \mathbb{R}^{m+1}$ is the concatenation of the decoder hidden state $\mathbf{s}_t$ and the context vector $\mathbf{c}_t$,$\mathbf{v}_p \in \mathbb{R}^m$ and $\mathbf{W}_p \in \mathbb{R}^{m\times(m+V)}$ are the parameters for the output layer.

The training target of our model is to reduce the Mean Square Error (MSE) between the predicted and the true values. The MSE is calculated as follows:

$$mse = \frac{1}{M} \sum_{i=1}^{M} (y'_i - y_i)^2, \tag{21}$$

where M is the length of target sequence.

IV. EXPERIMENTAL SETTINGS

Firstly, we provide the details of the datasets. Then, we explain the evaluation metrics for stock price prediction. Lastly, we present the comparative models and the set-up of our model.

A. DATA COLLECTION

Three datasets with news and correlated numerical data are collected from China Securities Index 300 (CSI 300) and Standard and Poor's 500 (S&P 500).

TABLE 1. Dataset partition.

Dataset	Training		Validation		Testing	
	Begin	End	Begin	End	Begin	End
D-CSI300	02/07/2001	12/31/2012	01/01/2013	06/30/2013	07/01/2013	08/04/2015
D-S&P500	10/20/2006	12/31/2012	01/01/2013	03/30/2013	04/01/2013	11/04/2013
M-CSI300	01/01/2016	09/30/2016	10/01/2016	10/30/2016	11/01/2016	12/30/2016

1) MINUTELY CSI300 (M-CSI300)

This dataset contains high-frequency data at the minute level. We develop a web crawler to collect news from five major financial news websites, which consist of.[1] It includes 780,920 financial news headlines. The numerical data is obtained from the SINA finance API.[2] The dataset holds the high-frequency financial data from Jan.1, 2016 to Dec.31, 2016. It provides intraday transaction information including price, volume and time in the minute-level.

2) DAILY CSI300 (D-CSI300)

In this dataset, we collect 24 million news between Feb. 07, 2001 and Dec. 30, 2015 (15 years) from SINA.[3] One of the major websites in China [21]. The numerical data of the CSI300 companies in the same period are downloaded from Yahoo Finance. There are nearly 969,600 of numerical data (3232×300).

3) DAILY S&P500 (D-S&P500)

We use the news dataset released by Ding *et al.* [11] 2015. It owns almost 550,000 financial news from 2006-10-20 to 2013-11-30. The news comes from Bloomberg and Reuters. In addition, we download numerical data of S&P500 companies in the same period from the Yahoo Finance. There are almost 881,000 numerical data points (1762×500).

Each dataset is split into the training, validation and testing set. The date information of each subset is recorded in Table. 1.

4) PREPROCESSING

The news without company names or their codes in headlines are filtered. Firstly, the headlines are tokenized and normalized. Secondly, headlines are aligned and labeled.

Tokenization. Although existing tokenizer produce outputs with high quality [31], [43], many financial terminologies cannot be identified correctly. A finance dictionary is thus employed to refine the segmentation results.

Normalization. This replaces the company name/code with "COMPANY," and the numbers with "NUMBER" [9].

[1]Five major financial news websites in china:

- Sina Finance (http://finance.sina.com.cn/)
- NetEase Finance (http://money.163.com/)
- SOHU Finance (http://business.sohu.com/)
- Tencent Finance (http://finance.qq.com/)
- Hexun (http://news.hexun.com)

[2]http://hq.sinajs.cn/. For instance, get the current information of sh600581: http://hq.sinajs.cn/?format=text&list=sh600581

[3]http://www.sina.com.cn

TABLE 2. Companies and standard deviation on testing set.

Dataset	Companies	Std
D-CSI300	202	18.78
D-S&P500	405	47.54
M-CSI300	185	37.84

For instance, a title in English before and after normalization is shown as follow:

Before: 3M sells drug unit for $2.1 billion in 3-part

After: COMPANY sells drug unit for NUMBER billion in 3-part deal

Alignment. The news and numerical data of a given company are aligned by news release time t. The news before t is reflected in the price changes. Thus, they can be ignored [38]. The window of length-K numerical data before news release time is chosen as the input sequence.

Labeling. The length-M close price after the news release time t is used as the labels. We employ $y_1, y_2, \cdots, y_M$ to denote the target sequence.

The statistics of each dataset is summarized in Table. 2. We filter out some of the companies without any news.

B. EVALUATION METRICS

We use Mean Squared Error (MSE) to evaluate the closeness between predicted and actual prices, and accuracy (ACC) to evaluate to which extend the predicted prices move in the same direction as the real prices [42].

The MSE of a predictor measures the average of the squares of the errors. If y_i' is the ith predicted value from a predictor, and y_i is the ith actual value, then the MSE of the predictor is computed as Eq. 21.

The accuracy is often used to evaluate the directional performance of a model. Given the stock price at i-th time step y_i and the stock price at news release time y_0, the trend is defined by

$$tr_i = \begin{cases} 1 & y_i > y_0 \\ 0 & y_i \leq y_0, \end{cases} \tag{22}$$

where $i \in [1, M]$ is the time steps in decoding. For each time step, we calculate the predicted trend tr_i' and the real trend tr_i:

$$acc_i = \frac{\sum_N I(tr_i' = tr_i)}{N} \tag{23}$$

where N is the total number of samples in dataset, $I(\cdot)$ is an indicator function.

TABLE 3. Results of stock price predction.

Methods	Representation	D-CSI300		D-S&P500		M-CSI300	
		MSE	Avg ACC	MSE	Avg ACC	MSE	Avg ACC
MKL	PV	0.611	0.516	0.171	0.508	1.004	0.601
ELM	PV	0.625	0.516	0.199	0.546	1.039	0.549
AZFin Text	BoW	0.646	0.516	0.162	0.539	1.039	0.539
AZFin Text-I	PV	0.612	0.517	0.184	0.512	1.001	0.593
DL4S	PV	0.622	0.505	0.158	0.556	0.993	0.526
Attention RNN	PV	0.624	0.510	0.158	0.560	0.999	0.519
NBAa	PV	**0.602**	0.518	**0.147**	0.573	**0.970**	0.601
NBAb	Word2vec	0.605	0.517	0.149	0.562	0.975	0.601
NBAc	Glove	0.609	**0.519**	0.150	0.576	0.988	0.605
NBAd	SDAE	0.606	**0.519**	0.148	**0.581**	0.981	**0.608**

Bold numbers indicate the best results

C. TRAINING SETUP

1) PARAMETERS

The textual representations (PV, Word2vec, Glove, and SDAE) are separately trained in news headlines from D-CSI300 and D-S&P500 under 100 dimensions. We employ **NBAa** - **NBAd** to denote these variations of our models with different textual representations. The numerical data contains the price information (open, high, low and close price) and trade volume information ($V = 5$). The length of the input sequence K and target sequence M of numerical data is set as $K = 30$ [36] and $M = 10$ [45] (In M-CSI300, we use the close price from $+20$ to $+29$ minutes as the targets [21]). We select the hidden units from [128, 64] for LSTM and attention on the validation set. The hidden unit of LSTM encoder and decoder are set to 64, and the attention hidden units is set to 64. During the training process, the Adam [19] optimization algorithm is used to minimize the Mean Squared Error (MSE) between the target price sequence and the predicted price sequence.

2) BASELINES

Four baseline models are adapted to predict the stock prices with the news and numerical data. In baseline 1 and 2, multiple models have trained under the direct strategy for multi-steps ahead prediction.

- AZFin Text [38]. The numerical data and headlines of news in Bag-of-Words (BoW) is fed into SVR to make stock price prediction .
- AZFin Text-I [38]. To get a fair comparison, we improve the AZFin text by encoding the headline into a dense representation via Paragraph Vector.
- DL4S [1]. We make some modifications on the original model. The numerical vector encoded by an LSTM concatenated with news embedding is fed into a multi-output layer to predict multiple targets.
- Attention RNN [36]. The numerical data is the input of sequence-to-sequence model with attention. In decoding, the context vector concatenated with news embedding is fed into a linear layer to predict the stock price.
- Multi-kernel Learning (MKL) [23] Multi-Kernel Learning uses a different kernel to process each source of information. It aims to transform different sources of information into the same feature space. Then, it categories samples into separated classes.
- Extreme Learning Machine (ELM) [27] ELM initializes the hidden weights randomly. It basically ensembles multiple simple classifiers to do the classification. Accordingly, it usually achieves faster training time than SVM.

V. RESULTS AND ANALYSIS

Firstly, the model's performance is compared with the baseline models on three datasets. Then, the significance of the proposed model in reducing the noise in stock price prediction is discussed. Lastly, the effects of the proposed model in utilizing the trend information are analyzed.

A. COMPARING TO BASELINES

Stock price prediction is a challenging task and a minor improvement usually leads to large potential profits [38]. To demonstrate the effectiveness of our NBA model, we compare it against the six baseline methods on three datasets. The results are listed in Table. 3. Examining the experimental results, we reach the following conclusions.

1) Our model achieves the best predictive performance in both MSE and accuracy. From the perspective of MSE, our model achieves the best performance on all three datasets. Especially, NBAa has 6.96% and 2.32% improvement compared to the best baseline models on D-S&P500 and M-CSI300. From the perspective of accuracy, our model also achieves the best results in these three datasets. Especially, NBAd rises the accuracy 2.32% and 1.35% higher than the best baseline models on D-S&P500 and M-CSI300.
2) Our NBA model can effectively capture the complementary between news and numerical data. First, given the same input, our NBA achieves better performance than AZFin Text-I, DL4S. We can conclude that this is due to the fact that NBA is more effective in capturing the complementary between dual sources. Second, given the same input and the same attention framework, our NBA achieves better performance than attention RNN. This is likely due to that our NBA transforms the news into the distribution of importance to select

numerical data. Moreover, the gate unit could filter noise in news. The concatenation of news embedding and context vector preserves the noise within news.

3) Our model is less sensitive to the change of news representation comparing to the SVR model. When using the PV representation to construct news embedding, the performance of SVR model is raised significantly on D-CSI300 and M-CSI300 and decreased on D-S&P500. Compared to AZFin Text and AZFin Text-I, our model has small changes in performance when using four different news representations. This is a benefit from the structure of the NBA model, which is mainly based on the numerical data.
4) Our model can effectively generate fusion features of news report and numerical data for stock trend prediction. We use the output of the penultimate layer of NBAa as the features and feed them into an SVM model to predict future trends. As shown in Table. 4, NBAa+SVM gets significant accuracy improvement. Especially, it yields a 6.04% improvement in accuracy on M-CSI300.
5) Trend predictions of the stock price at the minute level achieve a higher level of performance than those at day level. The prediction accuracy of models on M-CSI300 is significantly higher than those on D-CSI300 and D-S&P500. Since the news in minute level, which often contains only single piece of news, have less noise. The daily news often contains information which results in contradiction conclusions.

TABLE 4. The accuracy results of NBAa-SVM.

	D-CSI300	D-S&P500	M-CSI300
Accuracy	0.5281	0.5961	0.6447
Improvement(%)	+1.75	+2.60	+6.04

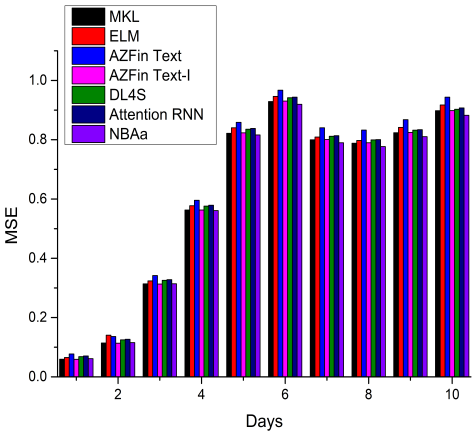


FIGURE 5. The MSE results on D-CSI300.

B. SIGNIFICANCE IN REDUCING NOISE

In order to show the significance of our model in reducing noise, we draw MSE results of baseline models and the NBAa on each dataset. We have seven MSE results for each time step in the next ten time steps. Fig. 5 illustrates the MSE results on D-CSI300. Fig. 6 shows the MSE results on D-S&P500. Fig. 7 draws the MSE results on M-CSI300. We observe that the MSE of NBAa is lower than other models on all times steps. As the time interval increases, the MSE of our model has a more significant reduction compared to other models. Table. 5 lists the t-test results on the MSE results between NBAa and other baseline models. The results indicate that our model achieves significantly lower MSE than other models.

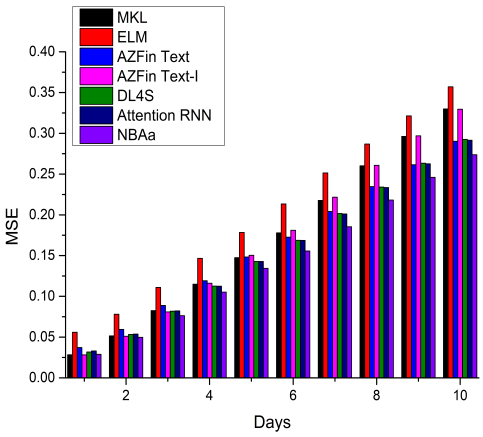


FIGURE 6. The MSE results on D-S&P500.

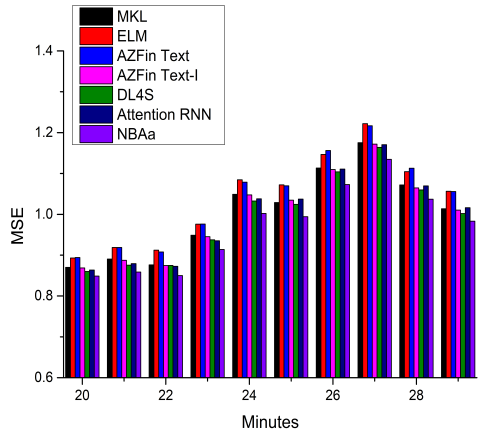


FIGURE 7. The MSE results on M-CSI300.

The improvement of MSE at each time step can show the effectiveness of Numerical-based attention in reducing noise. It can be observed that our model has significant lower MSE than baseline models on all three datasets. This confirms our assumption: The structure of NBA can improve the predict performance. This improvement benifits from tranforming news into the distribution of importance and filtering news information by a gate unit. The baseline models receive all the information in the news and numerical data directly. Thus, our model achieves the lowest MSE on all datasets.

C. EFFECTS IN UTILIZING TREND INFORMATION

To show the effects of utilizing trend information in news, we illustrate the accuracy of NBAa+SVM, NBAa, DS4L and Attention RNN on each dataset. These models share a similar

TABLE 5. The t-test results between the MSE results of different models.

	Dataset	AZFin Text	AZFin Text-I	DL4S	Attention RNN	MKL	ELM
NBAa($\times 10^{-3}$)	D-CSI300	**0.0086*****	**5.9532*****	**0.0035*****	**0.0012*****	**5.7267*****	**0.0262*****
NBAa($\times 10^{-3}$)	D-S&P500	**0.0001*****	**2.6397*****	**0.1532*****	**0.0737*****	**3.0478*****	**0.0089*****
NBAa($\times 10^{-5}$)	M-CSI300	**0.0012*****	**0.0226*****	**0.0626*****	**0.1539*****	**0.0070*****	**0.0015*****

p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

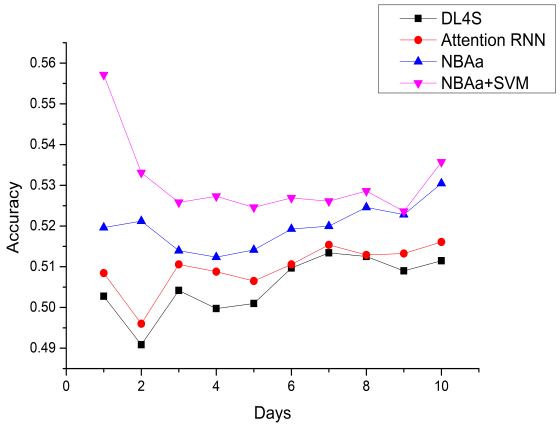


FIGURE 8. The average accuracy on D-CSI300.

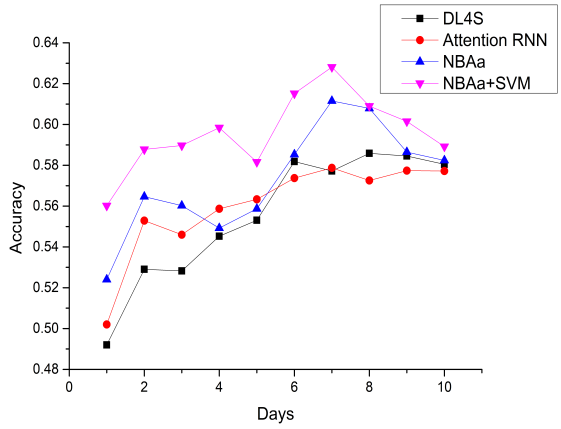


FIGURE 9. The average accuracy on D-S&P500.

model structure with NBAa. NBAa+SVM uses an extra SVM model compared to NBAa. DS4L uses fully connected layers to make predictions while NBA uses LSTM layers. Attention RNN uses the numerical information as attention guidance while NBAa uses news information.

Fig. 8, illustrates the accuracy results on D-CSI300. The accuracy of most models is slightly higher than 0.5. This is likely due to that the news come from a wide range of categories. News of certain categories may have a weak correlation to the stock market volatility. Fig. 9 shows the accuracy results on D-S&P500. The accuracy of the models increases over time. That may be caused by the skewness of data. Fig. 10 illustrates the accuracy results on M-CSI300. As shown in Fig. 8, the accuracy of models decreases over time in Fig. 10. In all figures, the accuracy of NBAa+SVM is significantly higher than that of other models, and the accuracy of NBAa is higher than other

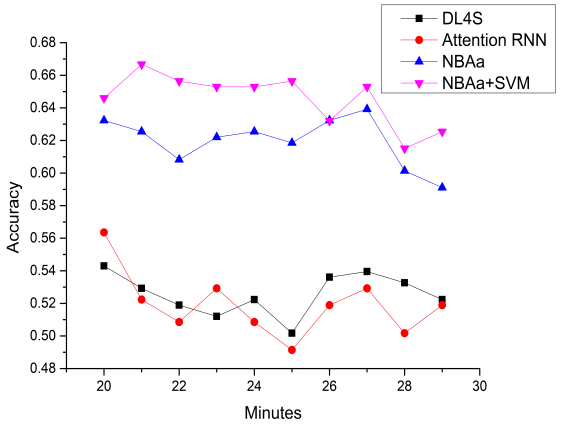


FIGURE 10. The average accuracy on M-CSI300.

TABLE 6. The t-test results between the MSE results of different models.

	Dataset	DL4S	Attention RNN
NBAa($\times 10^{-3}$)	D-CSI300	**0.0380*****	**0.4329*****
NBAa	D-S&P500	**0.0028*****	**0.0102****
NBAa($\times 10^{-6}$)	M-CSI300	**0.0753*****	**0.1834*****

p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

TABLE 7. The t-test results between the MSE results of different models.

	Dataset	DL4S	Attention RNN	NBAd
NBAa+SVM	D-CSI300	**0.0015*****	**0.0006*****	**0.0348****
NBAa+SVM($\times 10^{-4}$)	D-S&P500	**0.1116*****	**0.8470*****	**2.9818*****
NBAa+SVM($\times 10^{-6}$)	M-CSI300	**0.3498*****	**0.5159*****	**6.2333*****

p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

models in most cases. In order to show the significance of the differences between our model and the baseline models, we carry out the t-test between the accuracy of them. The test results are listed in Table. 7. The t-test results indicate that NBAa+SVM and NBAa have significantly higher accuracies than DL4S and Attention RNN on all datasets. And the accuracy of NBAa+SVM is significantly higher than NBAa. These results are consistent with our assumption that our model can effectively utilize the trend information within news by using them to select numerical data.

VI. CONCLUSIONS

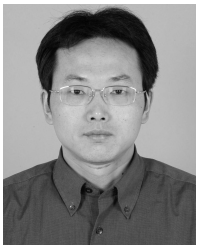
In this paper, we proposed a numerical-based attention (NBA) method to predict stock prices. In this method, the news is encoded to select the numerical data. Benefits from this transforming, noise is filtered and trend information of relevant stocks is utilized. In order to evaluate our method, three dual-source datasets source from the China Security Index 300 (CSI300) and Standard & Poor's 500 (S&P500) are build. Extensive experimental results on these three datasets

suggest that our NBA is superior to previous models in dual-source stock price prediction. The proposed method can effectively exploit the complementarity between news and numerical data in the stock market. In the future, we will explore the effectiveness of our NBA model in industrial or index level data.

REFERENCES

[1] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–6.
[2] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473
[3] J. Y. Campbell and L. Hentschel, "No news is good news: An asymmetric model of changing volatility in stock returns," *J. Financial Econ.*, vol. 31, no. 3, pp. 281–318, 1992.
[4] P.-C. Chang and C.-H. Liu, "A TSK type fuzzy rule based system for stock price prediction," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 135–144, 2008.
[5] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, and H. E. Stanley, "Which artificial intelligence algorithm better predicts the chinese stock market?" *IEEE Access*, vol. 6, pp. 48625–48633, 2018.
[6] W.-H. Chen, Y. Cai, K. Lai, and H. Xie, "A topic-based sentiment analysis model to predict stock market price movement using Weibo mood," *Web Intell.*, vol. 14, no. 4, pp. 287–300, 2016.
[7] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Syntax, Semantics Struct. Statist. Transl.*, 2014, p. 103.
[8] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Syst. Appl.*, vol. 83, pp. 187–205, Oct. 2017.
[9] K. Cortis *et al.*, "SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, 2017, pp. 519–535.
[10] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," in *Proc. IEEE 9th Int. Conf. Dependable, Autonomic Secure Comput.*, Dec. 2011, pp. 800–807.
[11] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. Ijcai*, 2015, pp. 2327–2333.
[12] G. Dong, K. Fataliyev, and L. Wang, "One-step and multi-step ahead stock prediction using backpropagation neural networks," in *Proc. IEEE 9th Int. Conf. Inf., Commun. Signal Process. (ICICS)*, Dec. 2013, pp. 1–5.
[13] Y. Guo, S. Han, C. Shen, Y. Li, X. Yin, and Y. Bai, "An adaptive SVR for high-frequency stock price forecasting," *IEEE Access*, vol. 6, pp. 11397–11404, 2018.
[14] E. Hadavandi, H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," *Knowl.-Based Syst.*, vol. 23, no. 8, pp. 800–808, 2010.
[15] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decis. Support Syst.*, vol. 55, no. 3, pp. 685–697, 2013.
[16] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," in *Proc. NAACL-HLT*, 2016, pp. 1367–1377.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[18] K.-J. Kim and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," *Expert Syst. Appl.*, vol. 19, no. 2, pp. 125–132, Aug. 2000.
[19] D. P. Kingma and J. Ba. (Dec. 2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[21] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements," *Inf. Sci.*, vol. 278, pp. 826–840, Sep. 2014.
[22] X. Li, Y. Rao, H. Xie, R. Y. K. Lau, J. Yin, and F. L. Wang, "Bootstrapping social emotion classification with semantically rich hybrid neural networks," *IEEE Trans. Affect. Comput.*, vol. 8, no. 4, pp. 428–442, Oct./Dec. 2017.
[23] X. Li, X. Huang, X. Deng, and S. Zhu, "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information," *Neurocomputing*, vol. 142, pp. 228–238, Oct. 2014.
[24] X. Li, C. Wang, J. Dong, F. Wang, X. Deng, and S. Zhu, "Improving stock market prediction by integrating both market news and stock prices," in *Proc. Int. Conf. Database Expert Syst. Appl.* Berlin, Germany: Springer, 2011, pp. 279–293.
[25] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, Oct. 2014.
[26] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. L. Wang, "Does summarization help stock prediction? A news impact analysis," *IEEE Intell. Syst.*, vol. 30, no. 3, pp. 26–34, May 2015.
[27] X. Li *et al.*, "Empirical analysis: Stock market prediction via extreme learning machine," *Neural Comput. Appl.*, vol. 27, no. 1, pp. 67–78, 2016.
[28] X. Li, H. Xie, T.-L. Wong, and F. L. Wang, "Market impact analysis via sentimental transfer learning," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 451–452.
[29] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[30] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, May 1970.
[31] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford coreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.
[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[33] A. K. Nassirtoussi, S. R. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 306–324, 2015.
[34] P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.
[35] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[36] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell. (Aug. 2017). "A dual-stage attention-based recurrent neural network for time series prediction." [Online]. Available: https://arxiv.org/abs/1704.02971
[37] Y. Rao, H. Xie, J. Li, F. Jin, F. L. Wang, and Q. Li, "Social emotion classification of short text via topic-level maximum entropy model," *Inf. Manage.*, vol. 53, no. 8, pp. 978–986, Dec. 2016.
[38] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, p. 12, 2009.
[39] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for long-term prediction of time series," *Neurocomputing*, vol. 70, nos. 16–18, pp. 2861–2869, Oct. 2007.
[40] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[41] X. Tang, C. Yang, and J. Zhou, "Stock price forecasting by combining news mining and time series analysis," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 1, Sep. 2009, pp. 279–282.
[42] P. Veronesi, "Stock market overreactions to bad news in good times: A rational expectations equilibrium model," *Rev. Financial Stud.*, vol. 12, no. 5, pp. 975–1007, 1999.
[43] G. Wu, D. He, K. Zhong, X. Zhou, and C. Yuan, "Leveraging rich linguistic features for cross-domain Chinese segmentation," in *Proc. 3rd CIPS-SIGHAN Joint Conf. Chin. Lang. Process.*, 2014, pp. 101–107.
[44] B. Xie, R. Passonneau, L. Wu, and G. G. Creamer, "Semantic frames to predict stock price movement," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 873–883.
[45] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 1970–1979.
[46] Y. Zhai, A. Hsu, and S. K. Halgamuge, "Combining news and technical indicators in daily stock price trends prediction," in *Proc. Int. Symp. Neural Netw.* Berlin, Germany: Springer, 2007, pp. 1087–1096.

GUANG LIU received the B.A. degree in business administration from Beihang University, Beijing, China, in 2008. He is currently pursuing the Ph.D. degree in intelligent science and technology with the Beijing University of Posts and Telecommunications, Beijing. His research interests include natural language processing in economics, deep learning, and application, reinforcement learning, QA system, and stock market analysis.



XIAOJIE WANG received the Ph.D. degree from Beihang University, in 1996. He is currently a Professor and the Director of the Centre for Intelligence Science and Technology, Beijing University of Posts and Telecommunications. His research interests include natural language processing and multi-modal cognitive computing. He is also an Executive Member of the Council of Chinese Association of Artificial Intelligence and the Director of Natural Language Processing Committee. He is also a member of the Council of Chinese Information Processing Society and the Chinese Processing Committee of China Computer Federation.