IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 29, 2018, accepted August 27, 2018, date of publication September 13, 2018, date of current version October 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2869735

Stock Market Prediction via Multi-Source Multiple Instance Learning

XI ZHANG[1], (Member, IEEE), SIYU QU[1], JIEYUN HUANG[1], BINXING FANG[1], AND PHILIP YU[2], (Fellow, IEEE)

[1]Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2]Department of Computer Science, The University of Illinois at Chicago, Chicago, IL 60607, USA

Corresponding author: Xi Zhang (zhangx@bupt.edu.cn)

This work was supported in part by the State Key Development Program of Basic Research of China under Grant 2013CB329605, in part by the Natural Science Foundation of China under Grant 61300014, in part by the NSF under Grant IIS-1526499, Grant IIS-1763325, and Grant CNS-1626432, and in part by the DongGuan Innovative Research Team Program under Grant 201636000100038.

ABSTRACT Forecasting the stock market movements is an important and challenging task. As the Web information grows, researchers begin to extract effective indicators (e.g., the events and sentiments) from the Web to facilitate the prediction. However, the indicators obtained in previous studies are usually based on only one data source and thus may not fully cover the factors that can affect the stock market movements. In this paper, to improve the prediction for stock market composite index movements, we exploit the consistencies among different data sources, and develop a multi-source multiple instance model that can effectively combine events, sentiments, as well as the quantitative data into a comprehensive framework. To effectively capture the news events, we successfully apply a novel event extraction and representation method. Evaluations on the data from the year 2015 and 2016 demonstrate the effectiveness of our model. In addition, our approach is able to automatically determine the importance of each data source and identify the crucial input information that is considered to drive the movements, making the predictions interpretable.

INDEX TERMS Stock prediction, multiple instance, event extraction, sentiment analysis.

I. INTRODUCTION

Stock markets play important roles in the economic operations of modern society. The estimation of the stock market index is of clear interest to various stakeholders in the market. According to the Efficient Market Hypothesis (EMH) [1], the stock market prices reflect all available information, and thus the prediction naturally relies on information from multiple sources, which can be roughly categorized into (1) quantitative data, e.g., historical prices, turnover rate, and (2) qualitative descriptions, such as the annual reports, announcements, news and social media posts. It is challenging to deal with qualitative data as they are usually unstructured and thus extracting useful signals from them is not trivial.

Along with the growing Web information and the advance of Natural Language Processing (NLP) techniques, recent works begin to explore Web news for market prediction. A number of existing studies have shown that the events reported in news are important signals that can drive market fluctuations [2]–[4]. However, most of the previous works represent news documents using simple features (e.g., bag-words, noun phrases, named entities) [5], [6], which may discard syntax information. Due to the large volume and diverse expressions of the events, how to represent them as useful features, and how to identify the crucial events that have significant impacts on the stock market are not trivial problems. In addition to events, a line of studies has shown that the investors' opinions can also largely influence the market volatility [7], [8]. With the prosperity of Web 2.0, the sentiments extracted from social media can be beneficial to predictions. Since both events and sentiments can drive the fluctuations of the market, it is natural to investigate how to effectively fuse them together to make a better prediction. The improvement may come from the correlations among different sources, and the consensus prediction with multi-source information can potentially outperform each prediction relying on a single source. This problem is analogous to the multi-labeler learning problem in crowdsourcing [9], [10],

but different from those studies that usually assume a labeler conducts classification with full information, each ''labeler'' (i.e., classifier) in this study is source-specific and only provided with limited information from its own source, making the consensus among labelers even more challenging.

In this work, we aim to learn a predictive model for describing the fluctuations in the stock market index by utilizing various sources of data, involving the historical quantitative data, the social media and Web news. The essential features we extract include the event representations from news articles and the sentiments from social media. Firstly, we propose a novel method to capture the event information. Specifically, structured events are extracted from news texts and then used as the inputs for Restricted Boltzmann Machines (RBMs) to do the pre-training. After that, the output vectors from RBMs are used as the inputs to a recently proposed sentence2vec framework [11], in order to achieve effective event embeddings. Secondly, we exploit the latent relationships among different data sources with carefully designed loss terms, and propose an extension of the Multiple Instance Learning (MIL) model that can effectively integrate the features from multiple sources to make more accurate predictions. One benefit of our method is that we can determine source-specific weights and identify the specific factors that incur the changes in the composite index. Figure 1 shows an example of the news precursors identified by our model, and the dots with numbers denote the probabilistic estimates for the events leading to the index change on Jan. 26, 2016.

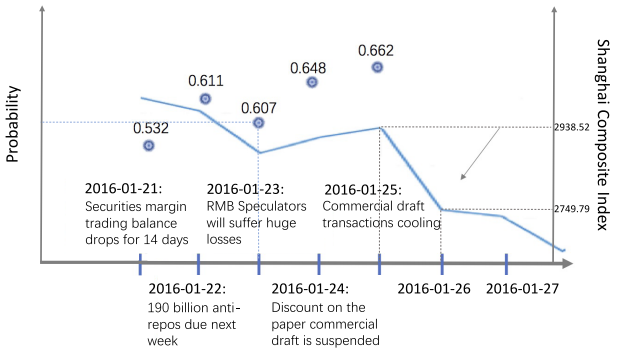

FIGURE 1. An example of the news events that are responsible for the Shanghai Composite Index change on Jan. 26, 2016. The x-axis is the timeline. The left y-axis is the probability of each event leading to the index change. The right y-axis is the composite index in Shanghai Stock Exchange.

The summary of the contributions is as follows:

1) To provide robust and accurate predictions for stock market movements, we extend the Multiple Instance Learning model to integrate the heterogeneous information including Web news, social media posts, and quantitative data.
2) The latent consistencies among different data sources are modeled in our framework by sharing the common estimated true label among the hinge losses of different data sources at the instance level.
3) A novel event representation model is proposed by first extracting structured events from news text, and then training them with deep learning methods involving RBM and sentence2vec to obtain dense vectors.
4) Evaluation results on two-year datasets show that our proposal can outperform the state-of-art baselines. Moreover, the impacts of different sources and the key factors that drive the movements can be obtained.

II. RELATED WORK

A. STOCK MARKET PREDICTION

There is a line of research works using event-driven stock prediction models. Hogenboom *et al.* [12] give an overview of event extraction methods. Akita *et al.* [13] convert newspaper articles into distributed representations via Paragraph Vector and model the temporal effects of past events with LSTM on opening prices of stocks in Tokyo Stock Exchange. Nguyen *et al.* [14] formulated a temporal sentiment index function, which is used to extract significant events. Then the corresponding blog posts are analyzed using topic modeling to understand the contents. Ding *et al.* [15] applied the Open IE tool to extract structured events from texts, and this event extraction method is also implemented as a baseline and compared with our proposal. Ding *et al.* [16] then trained event embeddings with a neural tensor network and then used the deep convolutional neural network to model influences of events.

In addition to events, investors' emotions also have great impacts on the stock market index. Bollen *et al.* [17] revealed that the public moods derived from Twitter have impacts on stock indicators. Si *et al.* [3] proposed a technique to leverage topic based sentiments from Twitter to predict the stock market. Makrehchi *et al.* [18] assigned a positive or negative label for each tweet according to stock movements. The aggregate sentiment per day shows predictive power for stock market prediction. Topic-specific sentiments are learned in [19] to facilitate the stock prediction. However, this method is not suitable to short texts in social media.

The common limitation of the aforementioned methods is that they rely only on a single data source and thus may limit the predictive power. In [20], events and sentiments are integrated into a tensor framework together with firm-specific features (e.g., P/B, P/E), to model the joint impacts on the stock volatility. We also implement it as a baseline. However, it uses a simple event extraction method which may not fully capture sufficient event information.

B. MULTIPLE INSTANCE LEARNING

The multiple instance learning (MIL) paradigm is a form of weakly supervised learning. Training instances arranged in sets are called bags or groups. A label is provided for entire groups instead of individual instances. Negative groups don't contain any positive instances, while positive groups contain at least one positive instance [21]. Various applications and the comparisons of different methods in MIL

were given in [22]. The common MIL approach is used to predict the group-level label, Liu *et al*. [23], however, proposed an approach to identify the instance-level labels, especially the labels of key instances in groups based on K nearest neighbors (K-NN). Kotzias *et al*. [24] predicted the labels for sentences given labels for reviews, which can be used to detect sentiments. A multiple-instance multiple-label learning framework with deep neutral network formation is proposed in [25]. An event forecasting framework via the nested multiple instance learning is proposed in [26]. However, it only uses one data source and simple event features, which may not be sufficient in the stock market application. We have implemented this algorithm as a baseline for comparison.

III. MULTI-SOURCE MULTIPLE INSTANCE MODEL

In this section, we first state and formulate the problem, and then propose the multi-source multiple instance (M-MI) framework. Before going into details of our framework, we define some important notations as shown in Table 1.

TABLE 1. Notations in our model.

Notation	Definition
$\mathbb{S} = \{\mathcal{G}\}$	a set of n multi-source super groups
$\mathcal{G} = \{C_i\}, i \in \{1, \cdots, t\}$	multi-source super group: a set of t groups
$C_i = \{\mathcal{X}_i, d_i, s_i\}$	an element of $\mathcal{G}$, a multi-source group
$\mathcal{X}_i = \{\boldsymbol{x}_{ij}\}, j \in \{1, \cdots, n_i\}$	an element of G, n_i is the number of instances in group $\mathcal{X}_i$
$\boldsymbol{x}_{ij} \in R^{V\times 1}$	a V-dimensional vector of news, the j-th instance in $\mathcal{X}_i$
$p_{ij} \in [0, 1]$	the prob. of $\boldsymbol{x}_{ij}$ in group $\mathcal{X}_i$ in multi-source super group to be positive
$\boldsymbol{p}_{m-i} \in [0, 1]$	the prob. of news group i in multi-source super group to be positive
$\boldsymbol{d}_i \in R^{3\times 1}$	a 3-dimensional vector of stock market data on day i, (average price, market index change and turnover rate)
$\boldsymbol{s}_i \in R^{2\times 1}$	a 2-dimensional vector of sentiment on day i (positive and negative)
$\boldsymbol{p}_{d-i} \in [0, 1]$	the prob. of stock market data on day i to the multi-source super group label that is positive
$\boldsymbol{p}_{s-i} \in [0, 1]$	the prob. of sentiment on day i to the multi-source super group label to be positive
$P_i \in [0, 1]$	the prob. of multi-source information (i.e., news, quantitative data and sentiments) on day i to be positive
$\mathcal{P} \in [0, 1]$	the estimated prob. for a multi-source super group
$Y \in \{-1, +1\}$	label of multi-source super group

A. PROBLEM STATEMENT

Stock markets are impacted by various factors, such as the trading volume, news events and the investors' emotions. Thus, relying on a single data source may not be sufficient to make accurate predictions. The object of our study is to develop a multi-source data integration approach to predict the stock market trends. Specifically, given a collection of economic news, social network posts and historical trading data, we aim to forecast the stock market index movements. Moreover, we also try to obtain the impacts of each data source and identify the key factors that have decisive influences, which may be influential news, collective sentiments or some important quantitative index in the trading data. These key factors are supporting evidence for further analysis and can make our prediction interpretable.

Formally, according to Table 1, a news article j on day i is denoted as a V-dimensional vector $\boldsymbol{x}_{ij} \in R^{V\times 1}$ (please note that the process of representing a news article as a vector will be illustrated in the next section). In order to predict the stock market movement on day $t + k$, we assume that there are a group of news articles for each day i ($i < t$), which is denoted as $\mathcal{X}_i$, and thus $\mathcal{X}_i = \{\boldsymbol{x}_{ij}\}, j \in \{1, \cdots, n_i\}$. In addition to the news articles, the sentiment and quantitative indices on day i (denoted as s_i and d_i respectively) are also taken into account. Then the temporal ordered collection of news articles, sentiments and quantitative indices across t days can be represented as a multi-source super group, that is, $\mathcal{G} = \{C_i\}, i \in \{1, \cdots, t\}$, where $C_i = \{\mathcal{X}_i, d_i, s_i\}$. The change in the stock market movement on day $t + k$ can be denoted as $Y_{t+k} \in \{+1, -1\}$, where $+1$ denotes the index rise and -1 denotes the index decline. Then the forecasting problem can be modeled as a mathematical function $f(\mathcal{G}) \rightarrow Y_{t+k}$, indicating that we map the multi-source information to an indicator (i.e., label) k days in the future from the day t, where k is number of the lead days that we aim to forecast.

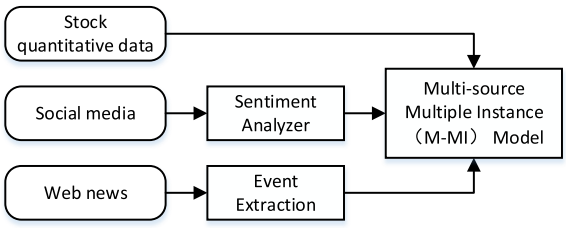

FIGURE 2. The system framework of our proposed model.

B. THE PROPOSED APPROACH

The framework of our proposal is shown in Fig. 2. The inputs of the framework are the stock quantitative data, the social media and Web news. We first use the sentiment analyzer to obtain the collective sentiments from social media, and extract effective event representations from the Web news. Then the extracted sentiments, events as well as the stock quantitative data are fed into the M-MI model. The M-MI model is proposed based on the Multiple Instance Learning algorithm, that is, a group of instances are given group labels, which are assumed to be an association function (e.g., OR, average) of the instance-level labels. Our work further distinguishes the instance-level labels, multi-source group-level labels, and multi-source super group-level labels. The primary goal is to predict the label for the multi-source super group that indicates the rise or decline of the stock market index. In addition, we also try to estimate the instance-level probabilities indicating how related a specific instance is to the index movement (i.e., target label), as well as the

source-specific weights that reveal how related a specific source is to the index movement.

To this end, for a given day, we first model the instance-level probability p_{ij} for a news article j on day i to the target label with a logistic function, that is

$$p_{ij} = \sigma(\boldsymbol{w_m}^T \boldsymbol{x}_{ij}) = \frac{1}{1 + e^{-\boldsymbol{w_m}^T \boldsymbol{x}_{ij}}} \tag{1}$$

where $\boldsymbol{w}_m$ denotes the weight vector of the news articles. The higher the probability p_{ij}, the more related the article j is to the target label. The probability of all the news articles for a given day i can be computed as the average of probabilities of each news article, that is

$$\boldsymbol{p}_{m-i} = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{ij} \tag{2}$$

In addition to news articles, we also model the probability $\boldsymbol{p}_{d-i}$ for stock quantitative data and $\boldsymbol{p}_{s-i}$ for sentiments on day i, that is

$$\boldsymbol{p}_{d-i} = \sigma(\boldsymbol{w}_d^T \boldsymbol{d}_i) = \frac{1}{1 + e^{-\boldsymbol{w}_d^T \boldsymbol{d}_i}} \tag{3}$$

$$\boldsymbol{p}_{s-i} = \sigma(\boldsymbol{w}_s^T \boldsymbol{s}_i) = \frac{1}{1 + e^{-\boldsymbol{w}_s^T \boldsymbol{s}_i}} \tag{4}$$

where $\boldsymbol{w}_d$ and $\boldsymbol{w}_s$ denote the weight vector of $\boldsymbol{d}_i$ and $\boldsymbol{s}_i$ respectively. We then model the probability $\boldsymbol{P_i}$ for multi-source information on day i as

$$\boldsymbol{P_i} = \theta_0 \boldsymbol{p}_{m-i} + \theta_1 \boldsymbol{p}_{d-i} + \theta_2 \boldsymbol{p}_{s-i} = \boldsymbol{\theta}(\boldsymbol{p}_{m-i}, \boldsymbol{p}_{d-i}, \boldsymbol{p}_{s-i})^T \tag{5}$$

where θ_0, θ_1 and θ_2 denote the source-specific weights of $\boldsymbol{p}_{m-i}$, $\boldsymbol{p}_{d-i}$ and $\boldsymbol{p}_{s-i}$ respectively, and $\theta_0 + \theta_1 + \theta_2 = 1$. It is obvious that $\boldsymbol{P_i} \in [0, 1]$. We use $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$ to denote the source weight vector, and then model the probability of the multi-source super group as the average of the probabilities in t days, that is

$$\mathcal{P} = \frac{1}{t} \sum_{i=1}^{t} \boldsymbol{P_i} \tag{6}$$

Then, we start with a log-likelihood loss function:

$$\min_{\boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}} \frac{1}{n} \sum_{\mathcal{G} \in \mathbb{S}} f(\mathcal{G}, Y, \boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta})$$
$$= \frac{1}{n} \sum_{\mathcal{G} \in \mathbb{S}} (-\mathbb{I}(Y = 1) log \mathcal{P} - \mathbb{I}(Y = -1) log(1 - \mathcal{P})) \tag{7}$$

where $\mathcal{G}$ is a multi-source super group, n is the number of multi-source super groups, and Y denotes the set of true labels. $\mathbb{I}(\cdot)$ is the indicator function.

As the influences of the multi-source information usually last for a number of days, we assume the probabilities on two consecutive days are essentially similar, which can be represented by minimizing the cost

$$g(C_i, C_{i-1}) = (\boldsymbol{P_i} - \boldsymbol{P_{i-1}})^2 \tag{8}$$

where C_i denotes the multi-source group for the day i. By introducing this loss term, Eq. 7 can be rewritten as:

$$\min_{\boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}} \frac{\beta}{n} \sum_{\mathcal{G} \in \mathbb{S}} f(\mathcal{G}, Y, \boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta})$$
$$+ \frac{1}{n} \sum_{C_i, C_{i-1} \in \mathcal{G}; \mathcal{G} \in \mathbb{S};} \frac{1}{t} \sum_{i=1}^{t} g(C_i, C_{i-1}, \boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}) \tag{9}$$

where β is a constant to control the contribution of the first term. Eq. 9 aggregates the costs at the super group level and the group level. However, the instance-level loss has not been considered yet, which is challenging to be designed due to two reasons: (1) it lacks of true labels at the instance level; (2) the instances from different sources are heterogeneous but intrinsically correlated. The instances can be categorized into three types according to their sources, and each type leads to a distinct loss term. Inspired by the hinge loss used in Support Vector Machines (SVMs), the classification loss term for the instances of news article instance x_{ij} is

$$h_1(\boldsymbol{x}_{ij}, \boldsymbol{w}_m) = max(0, m_0 - sgn(\boldsymbol{P_i} - \boldsymbol{P_0}) \boldsymbol{w}_m^T \boldsymbol{x}_{ij}) \tag{10}$$

Here, $sgn(\cdot)$ is the sign function, m_0 is a margin parameter used to separate the positive and negative instances from the hyperplane in the feature space. $\boldsymbol{w}_m^T \boldsymbol{x}_{ij}$ denotes the prediction with article x_{ij}. As the true label for each instance is unknown during the classifier training, we replace it with the estimated true label $sgn(\boldsymbol{P_i} - \boldsymbol{P_0})$, where $\boldsymbol{P_0}$ is a threshold parameter to determine the positiveness of the prediction. If $(\boldsymbol{P_i} - \boldsymbol{P_0}) > 0$, the prediction with multiple-source information on day i would be positive. Otherwise, it would be negative. Similarly, we can derive the instance-level loss terms for quantitative data and sentiments respectively,

$$h_2(\boldsymbol{d}_i, \boldsymbol{w}_d) = max(0, m_1 - sgn(\boldsymbol{P_i} - \boldsymbol{P_0}) \boldsymbol{w}_d^T \boldsymbol{d}_i) \tag{11}$$

$$h_3(\boldsymbol{s}_i, \boldsymbol{w}_s) = max(0, m_2 - sgn(\boldsymbol{P_i} - \boldsymbol{P_0}) \boldsymbol{w}_s^T \boldsymbol{s}_i) \tag{12}$$

Based on Eq. 10, 11 and 12, the classification loss at the instance level for each data source has been obtained. We then explain why they share a common estimated true label (i.e., $sgn(\boldsymbol{P_i} - \boldsymbol{P_0})$). As predictions from different sources are commonly correlated with each other, instead of treating the loss of each source independently, we need to consider their intrinsic consistencies. The intuition behind is that according to Efficient Market Hypothesis, different data sources would keep up to date with the latest stock market information, and they commonly indicate the same sign (index rise or fall). Thus, through sharing the same estimated true label, we are able to combine the indications from different data sources to learn a consensus label. This can potentially provide more robust and confident predictions.

We then give several cases to illustrate the consensus among sources. The three source-specific predictions are denoted as $\boldsymbol{l_0}$, $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ respectively. Firstly, if $\boldsymbol{l_0}$, $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ all make very positive predictions, i.e., large values of $\boldsymbol{p_{m-i}}$, $\boldsymbol{p_{d-i}}$ and $\boldsymbol{p_{s-i}}$, it would be confident to make a positive group-level prediction due to $\boldsymbol{P_i} > \boldsymbol{P_0}$ and $\boldsymbol{l_0}$, $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$

all agree with the label without costs. Secondly, if only $\boldsymbol{l_0}$ disagrees with the estimated true label, $\boldsymbol{l_0}$ will be penalized as $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ agree with this label and make $\boldsymbol{P_i}$ approach their predictions. Thirdly, if $\boldsymbol{l_0}$ disagrees with $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$, but $\boldsymbol{l_0}$ is very confident (and thus far from hyperplane) while $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ are not confident enough (and thus close to hyperplane), this may make $\boldsymbol{P_i}$ approach $\boldsymbol{l_0}$, resulting in that $\boldsymbol{l_0}$ agrees with the estimated true label while $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ disagree with it and thus are penalized. Our proposed instance-level loss terms are consistent with these cases and thus make sense.

Then we try to minimize the overall instance-level loss, that is, $h_1(\boldsymbol{x}_{ij}, \boldsymbol{w}_m) + h_2(\boldsymbol{d}_i, \boldsymbol{w}_d) + h_3(\boldsymbol{s}_i, \boldsymbol{w}_s)$. By introducing this summation and other regularization terms, the objective function Eq. 9 can be reformulated as

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}) \\
&= \frac{\beta}{n} \sum_{\mathcal{G}\in\mathbb{S}} f(\mathcal{G}, Y, \boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}) \\
&\quad + \frac{1}{n} \sum_{C_i, C_{i-1}\in\mathcal{G};\mathcal{G}\in\mathbb{S};} \frac{1}{t} \sum_{i=1}^{t} g(C_i, C_{i-1}, \boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta}) \\
&\quad + \frac{1}{n} \sum_{\boldsymbol{x}_{ij}\in\mathcal{X}_i;\mathcal{X}_i\in\mathcal{G};\mathcal{G}\in\mathbb{S}} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n_i} \sum_{j=1}^{n_i} h_1(\boldsymbol{x}_{ij}, \boldsymbol{w}_m) \\
&\quad + \frac{1}{n} \sum_{d_i, s_i\in C_i;C_i\in\mathcal{G};\mathcal{G}\in\mathbb{S}} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n_i} (h_2(\boldsymbol{d}_i, \boldsymbol{w}_d) + h_3(\boldsymbol{s}_i, \boldsymbol{w}_s)) \\
&\quad + \lambda_m R(\boldsymbol{w}_m) + \lambda_d R(\boldsymbol{w}_d) + \lambda_s R(\boldsymbol{w}_s) + \lambda_\theta R(\boldsymbol{\theta}) \qquad (13)
\end{aligned}$$

Eq. 13 is the ultimate objective function to optimize. To summarize, it consists of losses at three levels: the super group level, the group level and the instance level. In addition, it includes the regularization terms, that is, $R(\boldsymbol{w}_m)$, $R(\boldsymbol{w}_d)$, $R(\boldsymbol{w}_s)$ and $R(\boldsymbol{\theta})$, and β, λ_m, λ_d, λ_s and λ_θ are constants to control the trade-offs among multiple terms. The model learning goal is to estimate the parameters $\boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s$ and $\boldsymbol{\theta}$ to minimize $\mathcal{L}(\boldsymbol{w}_m, \boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{\theta})$. We randomly choose a set $(\mathcal{G}, Y)$ from $\mathbb{S}$, and the online stochastic gradient descent optimization is adopted to fit the model.

C. IDENTIFYING THE KEY FACTORS

After the learning process, the source weight vector $\boldsymbol{\theta}$ is obtained, representing the impacts of different data sources on the market movements. In addition, a probability for each piece of input information can also be obtained through Eq. 1, 3 or 4, which reveals the probability of that information signifying the rise of the market index on the target day. Note that if the probability signifying the index rise is p_r, the probability indicating index decline would be $1 - p_r$. We can identify the key input information that triggers the market index movement, if the product of its probability value and its source-specific weight is above a given threshold τ.

IV. FEATURE EXTRACTION

The quantitative features are quite simple to extract, we just collect three indices and normalize each index to form $\boldsymbol{d}_i \in R^{3\times 1}$. Here we introduce how to extract event representations from news articles and extract the sentiments from posts in social media, which are used as the inputs to M-MI framework.

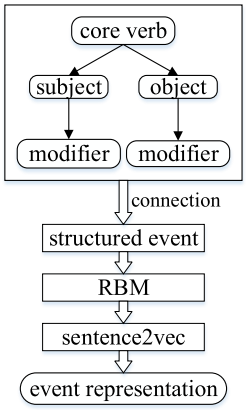

FIGURE 3. Structured event extraction from texts.

A. EVENT FEATURE EXTRACTION

Conventional methods commonly represent events using simple features such as TF-IDF, noun phrases or named entities. Recent advances in NLP techniques enable more accurate event models with structures. In this study, we first use the syntactic analysis method to extract the main structure information of the sentences, and then use it as the input to an RBM. The output of an RBM would be a pre-trained vector used as the input to sentence2vec, and then the event representations are obtained. The process is shown in Figure 3 and described in detail as follows. Note that though we use the Chinese dataset in this study, this process can also be applied to other languages.

1) **Structured event extraction.** With a commonly used text parser HanLP,[1] we can capture the syntactic structure of a sentence, which is depicted as a three-level tree at the top of Fig. 3. The root node denotes the core verb, and the nodes of the second layer are the subject of the verb and the object of the verb respectively. The child of the subject is the modifier who is the nearest to the subject in the sentence, and so is the child of the object. Then we connect these core words together as the structure information to represent the event information.
2) **Training with RBM.** We then map the structured event into a vector. To make the vectors better reconstruct the original events, we use RBM as a pre-training module. The Restricted Boltzmann Machine (RBM) is a generative stochastic artificial neural network, and has been applied in various applications such as dimensionality reduction [27]. RBM contains two-layer neural nets, one is the visible layer or input layer, and the

[1] https://github.com/hankcs/HanLP

other is the hidden layer. In our model, each event is represented as an m-dimensional vector with one-hot encoding, which is the visible layer. Our target is to estimate the n-dimensional hidden layer to approximate the input layer as much as possible. Then the hidden layer will be set as the initial vector in sentence2vec. The reason is that directly training the representations using sentence2vec without RBM may fall into the local minimum.

3) **Training with sentence2vec.** Finally, we use sentence2vec, the neural probabilistic language model to obtain the event representations. Different from the word2vec with CBOW model, the sentence id will be added during the training process of sentence2vec, and will also be mapped into a vector, called sentence vector, which would be the final vector that we want. In the training process, the sentence vector and the word vectors of context will be concatenated as the input to softmax. After training, the sentence vector will be obtained and used as the features for the proposed model.

Here is an example of extracting structured events from the news. The news text is that it is expected that the Renminbi speculators will face huge losses. After the dependency parsing analysis, the core words (Renminbi, speculators, face, huge, losses) are obtained, and after one hot coding, each word is encoded into zero or one vector. Then the vector preprocesses by RBM into a 100-dimensional vector, and finally processes by the sentence2vec became the news event feature vector. Through the above steps, the news event is obtained as a feature of the M-MI model, a 100-dimensional vector.

B. SENTIMENT EXTRACTION

To extract the sentiments from the posts in the social network, we use the LDA-S method [28], an extension of Latent Dirichlet Allocation (LDA) model that proposed to obtain the topic-specific sentiments for short texts. The intuition behind is that extracting sentiments discarding topics may be not sufficient as sentiment polarities usually depend on topics or domains [29]. In other words, the exact same word may express different sentiment polarities for different topics, e.g., the opinion word ''low'' in the phrase ''low speed'' in a traffic-related topic and ''low fat'' in a food-related topic. Therefore, extracting the sentiments corresponding to different topics can potentially improve the sentiment classification accuracy. The LDA-S model can infer sentiment distribution and topic distribution simultaneously for short texts. It consists of two steps. The first step aims to obtain the topic distribution of each post, and then set the topic as the one with the largest probability. The second step gets the sentiment distribution of each post.

In this work, a sentiment word list called NTUSD [30] is adopted, which contains 4370 negative words and 4566 positive words. If a word is an adjective but not in the sentiment word list, the sentiment label of this word is set as neutral. If a word is a noun, it is considered as a topic word. Otherwise, it is considered as a background word. For each topic, opinion word distributions are distinguished from two polarities, that is, positive or negative.

V. EXPERIMENTS

A. DATA COLLECTION AND DESCRIPTION

We collected stock market-related information from Jan. 1, 2015 to Dec. 31, 2016, and separate the information into two data sets, one for the year 2015 and the other for 2016. The data consist of three parts, the historical quantitative data, the news articles and the posts on the social network, which are introduced in detail as follows.

- **Quantitative data**: the source of quantitative data is Wind,[2] a widely used financial information service provider in China. The data we collect are the average prices, market index change and turnover rate of the Shanghai Composite Index in each trading day.
- **News data**: we collect the news articles on the macro economy through Wind, and get 38,727 and 39,465 news articles in 2015 and 2016 respectively. The news articles are aggregated by Wind from major financial news websites in China, such as http://finance.sina.com.cn and http://www.hexun.com. We process the news titles rather than the whole articles to extract the events, as the main topic of a news article is often summed up in the title.
- **Social media data**: the sentiments are extracted from the posts crawled from a popular investor social network in China named Xueqiu.[3] Totally 6,163,056 postings are collected for 2015 and 2016. For each post, we get the posting time stamp and the content.

For each trading day, if the stock market index rises, it would be a positive instance, otherwise it is a negative instance. For each year, we use the data from the first 10 months as the training set and the last 2 months as the testing set. We evaluate the performance of our model with varying lead days and varying historical days. Lead days refers to the number of days in advance the model makes predictions and the historical days indicates the number of days over which the multi-source information is utilized. The evaluation metrics we use are F1-score and accuracy (ACC).

B. COMPARISON METHODS

The following baselines and variations of our proposed model are implemented for comparisons. The full implementation of our framework is named as Multi-source Multiple Instance (M-MI) model.

- **SVM**: the standard support vector machine is used as a basic prediction method. During the training process, the label assigned to each instance and each group is the same as its multi-source super group label. During the

[2]http://www.wind.com.cn/
[3]https://xueqiu.com/

prediction phase, we obtain the predicted label for each of the instance, and then average the labels as the final label of the super group.

- **TeSIA:** the tensor-based learning approach [20] utilizes multi-source information for stock prediction. Specifically, it uses a third-order tensor to model the firm-mode, event-mode, and sentiment-mode data.
- **nMIL**: nested Multi-Instance Learning (nMIL) model [26] is the state-of-art baseline. In this model, only one data source, i.e., the news, is used to extract simple event features. It ignores the impacts of the sentiments and the historical quantitative indices.
- **O-MI**: Open IE Multiple Instance (O-MI) Learning model differs from M-MI in the event extraction module. It adopts a previously proposed event extraction method [15], and uses Open IE [31] to extract event tuples from sentences. The structured event tuples are then processed by sentence2vec to obtain event representations. Please note that the sentiment data and quantitative data are also used in this model.
- **WoR-MI**: Without RBM Multiple Instance (WoR-MI) Learning model is also a part of the M-MI framework. It differs M-MI in that it works without the RBM module, and therefore the sentence2vec module is fed with original structured events instead of pre-trained vectors.
- **WoH-MI**: Compare to M-MI, Without Hinge loss Multiple Instance (WoH-MI) Learning model lacks the instance-level hinge loss terms (i.e., Eq. 10, 11 and 12).

To make a fair comparison, we use the same set of instances and the same setting of parameters to evaluate different methods. In our proposal and the baselines, we set the predicted label to -1 if the estimated probability for a multi-source super group is less than 0.5; otherwise, we set the predicted label to $+1$.

TABLE 2. Prediction Results (history day=1, lead day=1).

Method	2015		2016	
	F-1	ACC	F-1	ACC
SVM	0.571	0.552	0.549	0.537
TeSIA	0.575	0.561	0.554	0.547
nMIL	0.581	0.563	0.557	0.541
O-MI	0.586	0.569	0.565	0.559
WoR-MI	0.592	0.577	0.584	0.568
WoH-MI	0.611	0.593	0.598	0.585
M-MI	**0.621**	**0.601**	**0.608**	**0.592**

C. PREDICTION RESULTS

We set both the number of history days and the number of lead days to 1. We empirically set $m_0 = 0.6$, and set m_1, m_2 and $\boldsymbol{P_0}$ all as 0.5, i.e., the default setting in hinge loss. β is set as 3.0, and λ_m, λ_d, λ_s and λ_θ are set as 0.05 by sensitivity analysis. The dimension of event representations is set as 100. Table 2 shows the performance of M-MI and the baselines. We can observe that M-MI outperforms all the baselines in both of the metrics, while SVM method shows the worst performance, indicating that simply tagging each news article with the label of its super-group is not effective. It can also be observed that M-MI and its variations (i.e., O-MI, WoH-MI and WoR-MI) all outperform nMIL. Compared to nMIL, M-MI improves F-1 by 6.9% in 2015 and 9.2% in 2016, while it improves accuracy by 6.7% and 9.4% in 2015 and 2016 respectively. Such gains mainly come from (1) utilizing multi-source information instead of only news articles, and (2) the advanced event representations rather than simple event features. Though all using multi-source information, TeSIA performs worse than M-MI and its variations, showing the effectiveness of our proposed models and feature extraction methods. Both M-MI and WoR-MI perform better than O-MI, indicating that both the structured event extraction module and the RBM pre-training module in our framework are effective. WoH-MI performs worse than M-MI, showing the proposed instance-level hinge losses across multiple data sources are useful for accurate predictions.

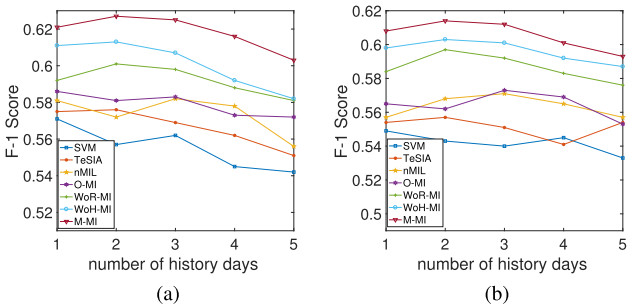

FIGURE 4. F-1 scores with varying history days. (a) 2015. (b) 2016.

Figure 4 (a) and (b) show the F-1 scores of all the comparative models with varying history days in training for 2015 and 2016 respectively (where lead day remains 1). The number of history days (i.e., t in Eq. 13) is varied from 1 to 5 and the results show that M-MI consistently performs better than the others. We can also observe that as the number of history days keeps increasing, the F-1 scores generally first go up and then go down. The possible reason is that the impacts of the news, sentiments and quantitative indices released on some day will quickly decay after a period of time (2 or 3 days). Thus, out-of-date information should be assigned with small weights or even discarded. Fortunately, our learning process can automatically assign small weights for information with weak impacts, alleviating the impact decaying problem.

In order to know how early our model can predict the index movement, we show the F-1 scores of WoR-MI and M-MI with varied lead days from 1 to 3 and history days from 1 to 5 in Table 3. We observe that as the number of lead days increases, the predictive capabilities of our models decrease. This makes sense since the stock market commonly reflects the available information in a timely manner. In other words, the up-to-date information will immediately be reflected in the index change and the impacts will decay as time goes, making it difficult for long-term predictions.

Figure 5 shows the weights of different data sources, that is, $\theta 1$, $\theta 2$ and $\theta 3$. It can be observed that among the

TABLE 3. F-1 scores for M-MI and WoR-MI in 2015 and 2016 with varying lead days.

		2015			2016		
	lead days	1	2	3	1	2	3
history day=1	WoR-MI	0.592	0.576	0.563	0.584	0.568	0.561
	M-MI	**0.621**	0.592	0.571	0.608	0.583	0.561
history day=2	WoR-MI	0.601	0.575	0.569	0.597	0.573	0.566
	M-MI	**0.627**	0.594	0.579	0.614	0.588	0.569
history day=3	WoR-MI	0.598	0.573	0.565	0.592	0.569	0.559
	M-MI	**0.625**	0.582	0.563	0.612	0.585	0.566
history day=4	WoR-MI	0.588	0.561	0.559	0.583	0.557	0.548
	M-MI	**0.616**	0.578	0.559	0.601	0.572	0.556
history day=5	WoR-MI	0.581	0.556	0.547	0.576	0.551	0.543
	M-MI	**0.603**	0.563	0.551	0.593	0.567	0.548

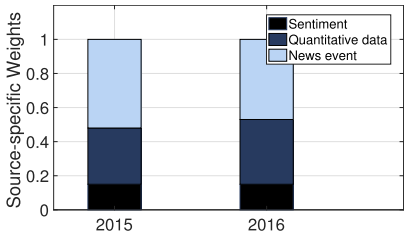

FIGURE 5. The weights of different data sources.

three sources, news events contribute most to the overall prediction, while the quantitative data takes the second place. It indicates that both news events and quantitative data have larger impacts to drive stock fluctuations than sentiments.

VI. CONCLUSIONS

In this paper, a Multi-source Multiple Instance model is proposed which can predict the stock market movement and identify the importance of the information simultaneously. Different from previous studies that commonly exploit only one data source, our model effectively integrates heterogeneous information, that is, the events, sentiments and historical quantitative features into a comprehensive framework, and considers the consistencies among different data sources to make a better prediction. We also propose a novel event representation learning process that can effectively capture the event information. Extensive evaluations on the two-year data confirm the effectiveness of our model.

REFERENCES

[1] E. F. Fama, "The behavior of stock-market prices," *J. Bus.*, vol. 38, no. 1, pp. 34–105, 1965.

[2] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," *Manage. Sci.*, vol. 53, no. 9, pp. 1375–1388, 2007.

[3] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based twitter sentiment for stock prediction," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2013, pp. 24–29.

[4] W. Y. Wang and Z. Hua, "A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Jun. 2014, pp. 1155–1165.

[5] S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith, "Predicting risk from financial reports with regression," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2009, pp. 272–280.

[6] R. Luss and A. D'Aspremont, "Predicting abnormal returns from news using text classification," *Quant. Finance*, vol. 15, no. 6, pp. 999–1012, 2015.

[7] R. R. Prechter, *The Wave Principle of Human Social Behavior and the New Science of Socionomics*, vol. 1. Gainesville, GA, USA: New Classics Library, 1999.

[8] J. R. Nofsinger, "Social mood and financial economics," *J. Behav. Finance*, vol. 6, no. 3, pp. 144–160, 2005.

[9] J. Bi and X. Wang, "Learning classifiers from dual annotation ambiguity via a min–max framework," *Neurocomputing*, vol. 151, pp. 891–904, Mar. 2015.

[10] S. Xie, W. Fan, and P. S. Yu, "An iterative and re-weighting framework for rejection and uncertainty resolution in crowdsourcing," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 1107–1118.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1188–1196.

[12] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, "An overview of event extraction from text," in *Proc. Workshop Detection, Represent., Exploitation Events Semantic Web (DeRiVE), 10th Int. Semantic Web Conf. (ISWC)*, vol. 779, 2011, pp. 48–57.

[13] 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.

[14] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Event extraction using behaviors of sentiment signals and burst structure in social media," *Knowl. Inf. Syst.*, vol. 37, no. 2, pp. 279–304, 2013.

[15] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using structured events to predict stock price movement: An empirical investigation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1415–1425.

[16] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 2327–2333.

[17] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[18] M. Makrehchi, S. Shah, and W. Liao, "Stock prediction using event-based sentiment analysis," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, vol. 1, Nov. 2013, pp. 337–342.

[19] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2015, pp. 1354–1364.

[20] Q. Li, L. Jiang, P. Li, and H. Chen, "Tensor-based learning for predicting stock movements," in *Proc. 29th AAAI Conf. Artif. Intell. (AAAI)*, 2015, pp. 1784–1790.

[21] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.

[22] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[23] G. Liu, J. Wu, and Z.-H. Zhou, "Key instance detection in multi-instance learning," in *Proc. Asian Conf. Mach. Learn.*, 2012, pp. 253–268.

[24] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 597–606.

[25] J. Feng and Z.-H. Zhou, "Deep MIML network," in *Proc. 21st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 1884–1890.

[26] Y. Ning, S. Muthiah, H. Rangwala, and N. Ramakrishnan, "Modeling precursors for event forecasting via nested multi-instance learning," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1095–1104.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[28] X. Zhang *et al.*, "IAD: Interaction-aware diffusion framework in social networks," *IEEE Trans. Knowl. Data Eng.*, to be published.

[29] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 56–65.

[30] L.-W. Ku and H.-H. Chen, "Mining opinions from the Web: Beyond relevance retrieval," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 12, pp. 1838–1850, 2007.

[31] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1535–1545.

XI ZHANG (M'17) received the Ph.D. degree in computer science from Tsinghua University. He was a Visiting Scholar at The University of Illinois at Chicago. He is currently an Associate Professor with the Beijing University of Posts and Telecommunications and is also the Vice Director of the Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, China. His research interests include data mining and computer architecture.

SIYU QU received the bachelor's degree in computer science from Xidian University in 2012. She is currently pursuing the master's degree with the Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Ministry of Education, China. Her research interests include data mining and machine learning.

JIEYUN HUANG received the bachelor's degree in information security from the Beijing University of Posts and Telecommunications in 2017, where she is currently pursuing the master's degree with the Key Laboratory of Trustworthy Distributed Computing and Service. Her research interests are in data mining and machine learning.

BINXING FANG received the Ph.D. degree from the Harbin Institute of Technology, China, in 1989. He was the Chief Scientist of the State Key Development Program of Basic Research of China. He is currently a member of the Chinese Academy of Engineering and is also a Professor with the School of Cyberspace Security, Beijing University of Posts and Telecommunications. His current research interests include big data and cybersecurity.

PHILIP YU (F'93) received the Ph.D. degree in electrical engineering from Stanford University. He is currently a Distinguished Professor in computer science at The University of Illinois at Chicago and is also the Wexler Chair in information technology. His research interests include big data, data mining, data stream, database, and privacy. He is a fellow of ACM. He received the Research Contributions Award from the IEEE International Conference on Data Mining in 2003, the Technical Achievement Award from the IEEE Computer Society in 2013, and the ACM SIGKDD 2016 Innovation Award. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and the *ACM Transactions on Knowledge Discovery from Data*.