Deep Learning for Event-Driven Stock Prediction

Xiao Ding†∗, Yue Zhang‡, Ting Liu†, Junwen Duan‡
†Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{xding, tliu, jwduan}@ir.hit.edu.cn
‡Singapore University of Technology and Design
yue_zhang@sutd.edu.sg

Abstract
We propose a deep learning method for event-driven stock market prediction. First, events are extracted from news text, and represented as dense vectors, trained using a novel neural tensor network. Second, a deep convolutional neural network is used to model both short-term and long-term influences of events on stock price movements. Experimental results show that our model can achieve nearly 6% improvements on S&P 500 index prediction and individual stock prediction, respectively, compared to state-of-the-art baseline methods. In addition, market simulation results show that our system is more capable of making profits than previously reported systems trained on S&P 500 stock historical data.

1 Introduction
It has been shown that the financial market is “informationally efficient” [Fama, 1965] — stock prices reflect all known information, and the price movement is in response to news or events. As web information grows, recent work has applied Natural Language Processing (NLP) techniques to explore financial news for predicting market volatility.

Pioneering work mainly uses simple features from news documents, such as bags-of-words, noun phrases, and named entities [Kogan et al., 2009; Schumaker and Chen, 2009]. Although useful, these features do not capture structured relations, which limits their potentials. For example, representing the event “Microsoft sues Barnes & Noble.” using term-level features {“Microsoft”, “sues”, “Barnes”, “Noble”} alone, it can be difficult to accurately predict the price movements of Microsoft Inc. and Barnes & Noble Inc., respectively, as the unstructured terms cannot differentiate the accuser (“Microsoft”) and defendant (“Barnes & Noble”).

Recent advances in computing power and NLP technology enables more accurate models of events with structures. Using open information extraction (Open IE) to obtain structured events representations, we find that the actor and object of events can be better captured [Ding et al., 2014]. For example, a structured representation of the event above can be (Actor = Microsoft, Action = sues, Object = Barnes & Noble). They report improvements on stock market prediction using their structured representation instead of words as features.

One disadvantage of structured representations of events is that they lead to increased sparsity, which potentially limits the predictive power. We propose to address this issue by representing structured events using event embeddings, which are dense vectors. Embeddings are trained such that similar events have similar vectors, even if they do not share common words. In theory, embeddings are appropriate for achieving good results with a density estimator (e.g. convolutional neural network), which can misbehave in high dimensions [Bengio et al., 2005]. We train event embeddings using a novel neural tensor network (NTN), which can learn the semantic compositionality over event arguments by combining them multiplicatively instead of only implicitly, as with standard neural networks.

For the predictive model, we propose to use deep learning [Bengio, 2009] to capture the influence of news events over a history that is longer than a day. Research shows diminishing effects of reported events on stock market volatility. For example, Xie et al. [2013], Tetlock et al. [2008] and Ding et al. [2014] show that the performance of daily prediction is better than weekly and monthly prediction. As shown in Figure 1, the influences of three actual events for Google

Figure 1: Example news influence of Google Inc.
We follow our previous work [Ding et al., 2014] and represent an event as a tuple \( E = (O_1, P, O_2, T) \), where \( P \) is the action, \( O_1 \) is the actor and \( O_2 \) is the object on which the action is performed. \( T \) is the timestamp of the event, which is mainly used for aligning stock data with news data, and not for event embeddings. For example, the event “Jan 13, 2014 - Google Acquires Smart Thermostat Maker Nest for $3.2 billion.” is modeled as: (Actor = Google, Action = acquires, Object = Nest, Time = Jan 13, 2014).

We extract structured events from free text using Open IE technology and dependency parsing. Given a sentence obtained from news text, we first use ReVerb [Fader et al., 2011] to extract the candidate tuples of the event \((O_1', P', O_2')\), and then parse the sentence with ZPar [Zhang and Clark, 2011] to extract the subject, object and predicate. We assume that \( O_1', O_2'\), and \( P' \) should contain the subject, object, and predicate, respectively. If this is not the case, the candidate tuple is filtered out. Redundancy in large news data allows this method to capture major events with high recalls.

2.2 Event Embedding

Events are extremely sparse. Our previous work used back-off features (e.g. \((O_1, P), (P, O_2), O_1, P, O_2\)) to address this issue, and we generalized verbs into verb classes, so that similar actions become one feature [Ding et al., 2014]. In contrast, our goal is to automatically learn embeddings for structured event tuples \( E = (O_1, P, O_2) \), which draw more fundamental relations between events, even if they do not share the same action, actor or object.

2 Neural Tensor Network for Learning Event Embeddings

2.1 Event Representation and Extraction

We follow our previous work [Ding et al., 2014] and represent an event as a tuple \( E = (O_1, P, O_2, T) \), where \( P \) is the action, \( O_1 \) is the actor and \( O_2 \) is the object on which the action is performed. \( T \) is the timestamp of the event, which is mainly used for aligning stock data with news data, and not for event embeddings. For example, the event “Jan 13, 2014 - Google Acquires Smart Thermostat Maker Nest for $3.2 billion.” is modeled as: (Actor = Google, Action = acquires, Object = Nest, Time = Jan 13, 2014).

We extract structured events from free text using Open IE technology and dependency parsing. Given a sentence obtained from news text, we first use ReVerb [Fader et al., 2011] to extract the candidate tuples of the event \((O_1', P', O_2')\), and then parse the sentence with ZPar [Zhang and Clark, 2011] to extract the subject, object and predicate. We assume that \( O_1', O_2'\), and \( P' \) should contain the subject, object, and predicate, respectively. If this is not the case, the candidate tuple is filtered out. Redundancy in large news data allows this method to capture major events with high recalls.
where $T_1^{[1:k]} \in \mathbb{R}^{d \times d \times k}$ is a tensor, and the bilinear tensor product $O_j^{[1:k]} P$ is a vector $r \in \mathbb{R}^k$, where each entry is computed by one slice of the tensor ($r_i = O_j^{[i]} P$, $i = 1, \ldots, k$). The other parameters are a standard feed-forward neural network, where $W \in \mathbb{R}^{k \times 2d}$ is the weight matrix, $b \in \mathbb{R}^k$ is the bias vector, and $f = tanh$ is the activation function. $R_2$ and $U$ are computed in the same way as $R_1$.

We also experiment with randomly initialized word vectors as the input of NTN, which is commonly used in related work on structured embedding [Bordes et al., 2011; Jenatton et al., 2012]. In our case, pre-trained word embeddings give slightly better results than randomly initialized embeddings.

### Training

We extract more than 10 million events from Reuters financial news and Bloomberg financial news as the training data for event embeddings. The training algorithm repeats for $N$ iterations over the training examples, which is a set of event tuples $E = (O_1, P, O_2)$, extracted from the training corpus using the method in Section 2.1. In each iteration, the training procedure is shown in Algorithm 1.

We assume that event tuples in the training set should be given a higher score than corrupted tuples, in which one of the event arguments is replaced with a random argument. The corrupted event tuple is denoted as $E' = (O_1', P, O_2)$. Specifically, we replace each word in $O_1$ with a random word $w'$ in our dictionary $D$ (it contains all the words in the training data) and derive a corrupted $O_1'$. We calculate the margin loss of the above two event tuples as:

$$
loss(E, E') = \max(0, 1 - f(E) + f(E')) + \lambda \| \Phi \|_2^2,
$$

where $\Phi = (T_1, T_2, T_3, W, b)$ is the set of parameters. The standard $L_2$ regularization weight $\lambda$ is set as 0.0001. If the loss $loss(E, E') = \max(0, 1 - f(E) + f(E'))$ is equal to zero, the algorithm continues to process the next event tuple. Otherwise, the parameters are updated to minimize the loss using the standard back-propagation (BP) algorithm. The iteration number $N$ is set to 500.

### Algorithm 1: Event Embedding Training Process

Input: $E = (E_1, E_2, \ldots, E_n)$ a set of event tuples; the model EELM

Output: updated model EELM'

1. random replace the event argument and got the corrupted event tuple
2. $E' \leftarrow (E_1', E_2', \ldots, E_n')$
3. while $E \neq []$
4. loss $\leftarrow \max(0, 1 - f(E) + f(E')) + \lambda \| \Phi \|_2^2$
5. if $loss > 0$
6. | Update(\Phi)
7. else
8. $E \leftarrow E \setminus \{E_i\}$
9. return EELM

Figure 3: Architecture of the prediction model based on a deep convolutional neural network.

### 3 Deep Prediction Model

We model long-term events as events over the past month, mid-term events as events over the past week, and short-term events as events on the past day of the stock price change. As shown in Figure 3, the prediction model learns the effect of these three different time spans on stock prices based on the framework of a CNN.

The input to the model is a sequence of event embeddings, where events are arranged in chronological order. Embeddings of the events on each day are averaged as a single input unit ($U$). The output of the model is a binary class, where Class +1 represents that the stock price will increase, and Class -1 represents that the stock price will decrease. As shown in Figure 3, for long-term (left) and mid-term (middle) news, the narrow convolution operation is used to combine $l = 3$ neighbour events. It can be viewed as feature extraction based on sliding window, which can capture local information through combinations of vectors in a window.

For our task, it is necessary to utilize all local features and predict stock price movements globally. Hence, we use a max pooling layer on top of the convolutional layer, which forces the network to retain only the most useful local features produced by the convolutional layer. Note that the convolution operation is only applied to the long-term and mid-term event embeddings, because the unit of timing is one day.

Formally, given a series of input event embeddings $U = (U_1, U_2, \ldots, U_n)$, where $U_j \in \mathbb{R}^d$, a one-dimensional convolution function takes the dot product of the weight vector $W_1 \in \mathbb{R}^d$ with each $l$ events (sliding window) in $U$ to obtain a new sequence $Q$:

$$
Q_j = W_1^T U_j - l + 1: j
$$

To determine the most representative features globally, we perform a max pooling operation over $Q$.

$$
V_j = \max Q(j, \cdot)
$$

where $Q(j, \cdot)$ is the $j$-th row of matrix $Q$. After max pooling, we obtain the feature vector $V$. For long-term and mid-term events, we obtain the feature vector $V^l$ and $V^m$, respectively.
For short-term events, we obtain the feature vector $V^s$ by directly using its averaged event embeddings $U^s$. The feature layer is the combination of long-term, mid-term and short-term feature vectors $V^C = (V^l, V^m, V^s)$.

To correlate the feature vector $V^C$ and stock prices, we use a feedforward neural network with one hidden layer and one output layer. Formally, let the values of the output layer be $y_{cls} (cls \in \{+1, -1\})$, its input be $net_{cls}$, and $Y$ be the neuron vector of the hidden layer; then: $y_{cls} = f(net_{cls}) = \sigma(W_2^T \cdot Y) \text{ and } Y = \sigma(W_3^T \cdot V^C)$, where $\sigma$ is the sigmoid function, $W_2$ is the weight vector between the hidden layer and the feature layer, and $W_3$ is the weight vector between the neuron $cls$ of the output layer and the hidden layer.

4 Experiments

4.1 Experimental Settings

We use financial news from Reuters and Bloomberg over the period from October 2006 to November 2013, released by Ding et al. [2014]1. Randinsky et al. [2012] and Ding et al. [2014] show that news titles are more useful for prediction compared to news contents. This paper extracts events only from news titles. We conduct our experiments on predicting individual stocks, obtaining indices and prices from Yahoo Finance. Detail statistics of training, development (tuning) and test sets are shown in Table 1.

Following Das and Chen [2007] and Xie et al. [2013], the standard measure of accuracy (Acc) and Matthews Correlation Coefficient (MCC) are used to evaluate S&P 500 index prediction and individual stock prediction. Following Lavrenko et al. [2000], we also evaluate the profitability of our proposed model.

<table>
<thead>
<tr>
<th>Training</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>#documents</td>
<td>442,933</td>
<td>110,733</td>
</tr>
<tr>
<td>#words</td>
<td>333,287,477</td>
<td>83,247,132</td>
</tr>
<tr>
<td>#events</td>
<td>295,791</td>
<td>34,868</td>
</tr>
<tr>
<td>time interval</td>
<td>02/10/2006 - 18/06/2012</td>
<td>19/06/2012 - 21/02/2013</td>
</tr>
</tbody>
</table>

Table 1: Statistics of datasets.

4.2 Baselines and Proposed Models

The baseline prediction systems are two-state-of-the-art financial-news-based stock market prediction systems: Luss and d’Aspremont et al. [2012] propose using bags-of-words to represent news documents, and constructing the prediction model by using Support Vector Machines (SVMs). Ding et al. [2014] report a system that uses structured event tuples $E = (O_1, P, O_2)$ to represent news documents, and investigates the complex hidden relationships between events and stock price movements by using a standard feedforward neural network.

In contrast to the baselines, we use a neural tensor network to learn event embeddings for representing news documents, and build a prediction model based on a deep CNN. To make detailed analysis, we construct the following five models:

* WB-NN: word embeddings input and standard neural network prediction model [Ding et al., 2014];
* WB-CNN: word embeddings input and convolutional neural network prediction model (this paper);
* E-CNN: structured events tuple [Ding et al., 2014] input and convolutional neural network prediction model (this paper);
* EB-NN: event embeddings input (this paper) and standard neural network prediction model [Ding et al., 2014];
* EB-CNN: event embeddings input and convolutional neural network prediction model (this paper).

A word embedding input (WB) consists of the sum of each word in a document; it addresses sparsity in word-based inputs, and can serve as a baseline embedding method. The standard feedforward neural network (NN) is used as a baseline to compare with the deep CNN.

4.3 Development Results

S&P 500 Index Prediction

We test the influence of event embeddings by comparing them with the structured event representation [Ding et al., 2014], and the CNN model by comparison with the standard feedforward neural network model [Ding et al., 2014]. The experimental results are shown in Table 2. We find that:

(1) Comparison between the word-embedding-based models and event-embedding-based models (e.g. WB-NN vs EB-NN and WB-CNN vs EB-CNN) confirms our previous conclusion [Ding et al., 2014]: events are better features than words for stock market prediction.

(2) Event embedding is useful for the task of stock market prediction. Given the same prediction model (CNN or NN), the event embeddings based methods (EB-NN and EB-CNN) achieve consistently better performance than the events-based methods (E-NN and E-CNN). This is likely due to the following reasons. First, low-dimensional dense vector can effectively alleviate the problem of feature sparsity. In fact, using word embeddings of events (WB-NN) only, we can achieve better performance than Ding et al. [2014] (E-NN). This contrast demonstrates the importance of reducing sparsity, which rivals the effect of structured information.

Second, we can learn deeper semantic relations between event embeddings, by modeling the semantic compositionality over word embeddings. For example, the two events $E_1 = (\text{Actor} = \text{Nvidia fourth quarter results}, \text{Action} = \text{miss}, \text{Object} = \text{views})$ in the training data and $E_2 = (\text{Actor} = \text{Delta profit},$...
Figure 4: Development results of individual stock prediction (companies are named by their ticker symbols).

| Action = didn’t reach, Object = estimates | in the development data result in different features by Ding et al. [2014] even after backing-off, as the actor and the object of $E_1$ and $E_2$ are different, and the predicates of these two events cannot be generalized to the same verb class. However, the semantic distance of these two event embeddings are very small in our model, even though they do not have similar word embeddings. As a result, $E_1$ can serve as a relative training example for predicting using $E_2$ in our model.

(3) CNN-based prediction models are more powerful than NN-based prediction models (e.g. WB-CNN vs WB-NN, EB-CNN vs EB-NN, and E-CNN vs E-NN). This is mainly because CNN can quantitatively analyze the influence of the history events over longer terms, and can extract the most representative feature vector for the prediction model.

**Individual Stock Prediction**

We compare our approach with the baselines on individual stock prediction using the development dataset. We use the 15 companies selected by Ding et al. [2014] from S&P 500. The list consists of samples from high-ranking, middle-ranking, and low-ranking companies from S&P 500 according to the Fortune Magazine. The results are shown in Figure 4 (as space is limited, we only show comparison between EB-CNN and the two baselines). We find that:

1. Our model achieves consistently better performance compared to the baseline methods, on both individual stock prediction and S&P 500 index prediction.
2. Our model achieves relatively higher improvements on those lower fortune ranking companies, for which fewer news are available. For the baseline methods, the prediction results of low-ranking companies dramatically decrease. However, our model considers the diminishing influence of monthly news and weekly news, which are important features for individual stock prediction. Hence, even without daily news, our model can also give relatively accurate prediction results.

**Market Simulation**

We simulate real stock trading by following the strategy proposed by Lavrenko et al. [2000], which mimics the behavior of a daily trader who uses our model in a simple way. If the model indicates that an individual stock price will increase the next day, the fictitious trader will invest in $10,000 worth of that stock at the opening price. After a purchase, the trader will hold the stock for one day. During the holding time, if the stock can make a profit of 2% or more, the trader sells immediately. Otherwise, at the end of the day, the trader sells the stock at the closing price. The same strategy is used for shorting, if the model indicates that an individual stock price will decrease. If the trader can buy the stock at a price 1% lower than shorted, he/she buys the stock to cover. Otherwise, the trader buys the stock at the closing price.

We use the same training, development and test dataset as shown in Table 1, for the simulation. Table 3 summarizes the average cumulative profits over the 15 companies in Section 4.3. These results are obtained on the development data. The cumulative earnings of our model averaged $16,785 (which means that trading $10,000 worth of stocks would result in a net profit of $16,785), which is higher than Luss and d’Aspremont [2012] ($8,694) and Ding et al. [2014] ($10,456). Except for the reasons analyzed in Sections 4.3, we notice that if there is no news reported for an individual stock on the previous day, their models cannot predict the trend of the stock price movements on a day, because they do not leverage long-term and mid-term news. This does not hurt the evaluation results of accuracy and MCC, but can hurt the real profit.

To verify the statistical significance of our earnings, we perform a randomization test [Edgington and Onghena, 2007] by randomly buying or shorting for 1000 trials. The mean profit over the randomized test is -$9,865 and the performance of our model is significant at the 1% level.

| Action = didn’t reach, Object = estimates | in the development data result in different features by Ding et al. [2014] even after backing-off, as the actor and the object of $E_1$ and $E_2$ are different, and the predicates of these two events cannot be generalized to the same verb class. However, the semantic distance of these two event embeddings are very small in our model, even though they do not have similar word embeddings. As a result, $E_1$ can serve as a relative training example for predicting using $E_2$ in our model.

(3) CNN-based prediction models are more powerful than NN-based prediction models (e.g. WB-CNN vs WB-NN, EB-CNN vs EB-NN, and E-CNN vs E-NN). This is mainly because CNN can quantitatively analyze the influence of the history events over longer terms, and can extract the most representative feature vector for the prediction model.

**Individual Stock Prediction**

We compare our approach with the baselines on individual stock prediction using the development dataset. We use the 15 companies selected by Ding et al. [2014] from S&P 500. The list consists of samples from high-ranking, middle-ranking, and low-ranking companies from S&P 500 according to the Fortune Magazine. The results are shown in Figure 4 (as space is limited, we only show comparison between EB-CNN and the two baselines). We find that:

1. Our model achieves consistently better performance compared to the baseline methods, on both individual stock prediction and S&P 500 index prediction.

<table>
<thead>
<tr>
<th>Average Profit</th>
<th>Luss and d’Aspremont [2012] $8,694</th>
<th>Ding et al. [2014] $10,456</th>
<th>EB-CNN $16,785</th>
</tr>
</thead>
</table>

Table 3: Averaged profit of 15 individual companies.

<table>
<thead>
<tr>
<th>Index Prediction</th>
<th>Acc</th>
<th>MCC</th>
<th>Individual Stock Prediction</th>
<th>Acc</th>
<th>MCC</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luss [2012]</td>
<td>56.38%</td>
<td>0.07</td>
<td>Ding [2014]</td>
<td>58.83%</td>
<td>0.16</td>
<td>EB-CNN</td>
</tr>
<tr>
<td></td>
<td>58.74%</td>
<td>0.25</td>
<td>$10,456</td>
<td>61.47%</td>
<td>0.31</td>
<td>$16,774</td>
</tr>
</tbody>
</table>

Table 4: Final results on the test dataset.
4.4 Final Results

Table 4 shows the final experimental results on the test dataset, where Luss [2012] is the model of Luss and d’Aspremont [2012] and Ding [2014] is the model of our previous work. As space is limited, we only show the average prediction results of 15 individual stocks. The results demonstrate consistently better performance, which indicates the robustness of our model.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Profit of Lavrenko et al. [2000]</th>
<th>Profit of EBCNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>$47,000</td>
<td>$42,000</td>
</tr>
<tr>
<td>Lucent</td>
<td>$20,000</td>
<td>$27,000</td>
</tr>
<tr>
<td>Yahoo</td>
<td>$19,000</td>
<td>$32,000</td>
</tr>
<tr>
<td>Amazon</td>
<td>$14,000</td>
<td>$35,000</td>
</tr>
<tr>
<td>Disney</td>
<td>-$53,000</td>
<td>$7,000</td>
</tr>
<tr>
<td>AOL</td>
<td>-$18,000</td>
<td>$14,000</td>
</tr>
<tr>
<td>Intel</td>
<td>-$14,000</td>
<td>$8,000</td>
</tr>
<tr>
<td>Oracle</td>
<td>-$13,000</td>
<td>$17,000</td>
</tr>
</tbody>
</table>

Table 5: Profit compared with Lavrenko et al. [2000]. (there are 4 negative profit stocks out of 15 which are not included in this table)

Efficient Market Hypothesis (EMH) [Fama, 1965] states that the price of a security reflects all of the information available, and that everyone has a certain degree of access to the information. Despite 50 years of studies from the fields of finance, computer science and other research communities, the debate continues over what kinds of information can be useful for stock market prediction. In Artificial Intelligence (AI), three sources of information has been the most exploited for algorithmic stock market prediction.

First, some prediction techniques leverage historical and time-series data [Taylor and Xu, 1997; Andersen and Bollerslev, 1997; Taylor, 2007]. Researchers believed that predictions can be made through careful averaging of historical price and volume movements and comparing them against current prices. It is also believed that there are certain high or low psychological price barriers, such as support and resistance levels. However, these methods ignore one key source of market volatility: financial news.

With advances of NLP techniques, various studies have found that financial news can dramatically affect the share price of a security [Cutler et al., 1998; Tetlock et al., 2008; Luss and d’Aspremont, 2012; Xie et al., 2013; Wang and Hua, 2014]. Recently, we proposed using structured events to represent news, which can indicate the actors and objects of events [Ding et al., 2014]. However, modeling complex event structures directly, their work is challenged by the new problem of sparsity. To this end, this paper proposes learning event embedding.

Apart from events, sentiment is another perspective of deep semantic analysis of news documents [Das and Chen, 2007; Tetlock, 2007; Tetlock et al., 2008; Bollen et al., 2011; Si et al., 2013]. Tetlock [2007] examines how qualitative information (i.e. the fraction of negative words in a particular news column) is incorporated in aggregate market valuations. Bollen and Zeng [2011] find that large-scale collective emotions (representing public moods) on Twitter is correlated with the volatility of Dow Jones Industrial Average (DJIA). Si et al. [2014] propose to regress topic-sentiment time-series and stock’s price time series. Their work is orthogonal to event-driven stock market prediction.

6 Conclusion

We demonstrated that deep learning is useful for event-driven stock price movement prediction by proposing a novel neural tensor network for learning event embeddings, and using a deep convolutional neural network to model the combined influence of long-term events and short-term events on stock price movements. Experimental results showed that event-embeddings-based document representations are better than discrete events-based methods, and deep convolutional neural network can capture longer-term influence of news event than standard feedforward neural network. In market simulation, a simple greedy strategy allowed our model to yield more profit compared with previous work.

Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the Na-


