A Tensor-based eLSTM Model to Predict Stock Price using Financial News

Jinghua Tan
Southwestern University of Finance and Economics
Jinghua_Tan@126.com

Jun Wang
Southwestern University of Finance and Economics
juujuu0058@163.com

Denisa Rinprasertmeechai
Southwestern University of Finance and Economics
meimi.denisa@gmail.come

Rong Xing
Southwestern University of Finance and Economics
337402185@qq.com

Qing Li
Southwestern University of Finance and Economics
liq_t@swufe.edu.cn

Abstract

Stock market prediction has attracted much attention from both academia and business. Both traditional finance and behavioral finance believe that market information affects stock movements. Typically, market information consists of fundamentals and news information. To study how information shapes stock markets, common strategies are to concatenate various information into one compound vector. However, such concatenating ignores the interlinks between fundamentals and news information. In addition, the fundamental data are continuous values sampled at fixed time intervals, while news information occurred randomly. Such heterogeneity leads to miss valuable information partially or twist the feature spaces. In this article, we propose a tensor-based event-LSTM (eLSTM) to solve these two challenges. In particular, we model the market information with tensors instead of concatenated vectors and balance the heterogeneity of different data types with event-driven mechanism in LSTM. Experiments performed on an entire year data of China Securities markets demonstrate the supreme of the proposed approach over the state-of-the-art algorithms including AZfinText, eMAQT, and TeSIA.

1. Introduction

A company's stock price reflects investor perception of its ability to earn and grow its profits in the future. The traditional Efficient Market Hypothesis (EMH) states that a stock price is always driven by unemotional investors [1][2]. New information related to markets will change investors' expectations on the markets and cause stock prices to move [3]. While behavioral finance studies attribute the stock movements to investors' cognitive and emotional bias. Although both theories have different view on how information shapes stock movements, both agree that the volatility of stock market comes from the release, dissemination and absorption of information [4].

In previous studies, scholars have found that stock movements are affected by various information sources including transaction data, news, social media, search behavior [5][6]. Some researchers take a further step by examining the joint effect of various types of information which is proved to be helpful to capture stock movements [7][8][9]. For example, Weng, Ahmed and Megahed applied transaction data, Google news and Wikipedia browsing information to predict stock markets and found that multi-dimensional data have better prediction performance. Schumaker and Chen quantified the financial news and achieved a better performance by integrating it with transaction data [8]. Essentially, stock markets are affected by multiple information sources which is roughly categorized into two subgroups, that is, the fundamentals and financial news [10].

The challenge lies in how to find the joint effect of fundamental data and news information on stock markets. Traditional strategies are to concatenate them into a super compound vector and utilize different models including Support Vector Machine, Decision Tree, and Artificial Neural Networks for predicting [11][12][13]. However, these vector-based models may fail to capture the interconnections of multiple information and ignore the inherent links. To achieve this, some scholars model the multi-dimensional information with tensors for better performance [14].

In addition, the fundamental data is characterized with continuous values sampled with fixed time intervals. In contrast, the news information are discrete values sampled with varying time intervals because of the randomness of news occurrence. In Figure 1, it shows the news event with stock "000001" between Jan 1, 2015 and Apr 30, 2015. It can be observed that the occurrence of news event is distributed irregularly with varying intervals ranging from days, weeks or

URI: https://hdl.handle.net/10125/59606
ISBN: 978-0-9981331-2-6

even months. How to fuse these two data types for a supervised learning problem is yet to be explored.

Previous researches typically solve this problem by using a part of training data, that is, only the data sampled at the event time are kept for further analysis. For instance, Schumaker and Chen utilized the transaction data within 20 minutes right after a breaking news is released to study media-aware stock movements [8]. In this way, only partial fundamental data are utilized and some valuable fundamental information is ignored. One possible solution is to apply the entire data with some missing values in the dimension of event. However, the sparsity in the event dimension will twist the entire feature space.

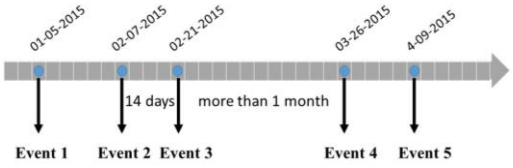


Figure 1. An example of the events with irregular time intervals.

To solve the above two challenges, we proposed a tensor-based event-LSTM (eLSTM) model with several unique contributions. Specifically, (1) to retain the interconnection of different information sources, we model market information with tensors instead of compound vectors. (2) We introduce an event-driven mechanism in Long-Short Term Memory (LSTM) to deal with the data fusion between continuous values sampled with fixed time intervals (fundamental data) and discrete values sampled with varying time intervals (sentiment data). Such mechanism is able to capture the special patterns hidden in the integrated data with sparseness (sparsity) in some dimensions. (3) Experiments performed on an entire year data of China Securities markets demonstrated the supreme of the proposed approach over the state-of-the-art algorithms including AZfinText, eMAQT, and TeSIA.

The remaining content of this article is organized as follows. We first describe related research in Section 2. The technical details of proposed model are presented in Section 3. We then test the model using real stock market data in Section 4. Section 5 concludes this paper with speculation on how the current prototype can be further improved.

2. Related Work

In this section, we review relevant literatures from two aspects, that is, the influence of information on stock volatility and the approaches to quantify such influence.

2.1 Information and stock volatility

Stock price reflects investors' expectations on a company's future cash flows. Investors may change their expectations as they receive new information, which cause stock market fluctuation. Stock market information can be roughly categorized into the fundamental information and media emotional information which extracted from financial news [10].

- **Fundamental information**: A number of studies in traditional finance have examined the effect of fundamental information such as firm size, cash flow, book-to-market equity, and historical data on stock movements [15][16][17]. For example, Bernard and Stober showed that cash flow could provide additional information content for better understanding of stock market. Fama and French found that a stock performance was mainly deter-mined by three risk factors including the overall market, the firm size, and the book-to-market equity ratio (BE/ME).
- **Media emotion information:** The studies of media-aware stock movements have started with the influence of financial reports on stocks markets. Later on, researchers also observed the influence of online media [12][13]. In particular, investors' decisions could be influenced by the opinions of others, which may result in a herd behavior in investment. For example, Schumaker and Chen experimented with several textual news representation approaches to study the effect of breaking news on stock movements [8]. Bollen, Mao and Zeng found that the collective mood states derived from 10 million tweets were correlated with Dow Jones Industrial Average (DJIA) using SOFNN model [12]. Three media-aware hedge funds including Derwent Capital Markets, DCM capital and Cayman Atlantic were established thereafter.

However, those studies utilized vector-based approaches that concatenate the fundamentals and news information into one compound vector. Such information linearization reduces or ignores the intrinsic association among various information sources. Some researchers take a further step by modeling market information with tensors to capture the interconnections of those different information sources [4].

2.2 Stock analysis models

Once fundamental information and news information are obtained in a machine-friendly form, various types of analysis models are proposed to study the media-aware stock movements. There are three mainstream models, i.e., statistical models in statistics, regression models in econometrics and machine learning models in computer science. Table 1 summarizes the related work in terms of these models.

Statistical models emphasize the correlation between stock markets and a single feature without considering others [5][6]. While economic models focus on the causal relationship between information sources and market movements and also ignore the interconnections of various sources [18][19]. For example, Huang, Teoh, and Zhang et al. applied logistic regression models to find that the abnormal optimism in a company earnings report dragged its stock performance.

With richness of information in social media, the influences of media information on stock markets become more complicated than ever before. Traditional statistical models and econometric models often fail to find complex non-linear relationship of multi-dimensional data [7]. Computer scientists take a further step by utilizing machine learning algorithms to capture the complex nonlinearity relationships between market information and stock movements [4].

Essentially, how market information affects stock markets is a binary classification problem. With the ability to take high-dimensional data as input, many machine learning algorithms are applied to solve this problem including SVM, Bayesian classifier and Decision Tree [11][13][20]. For example, one of the early researches can be traced back to the work of Wuthrich et al. which forecasted the daily trends of five major stock market indexes using neural network and KNN. Schumaker and Chen estimated a discrete stock price twenty minutes after a news article was released by using Support Vector Regression (SVR).

With the popularity of deep learning techniques, deep neural networks have succeeded in many fields including text processing [21], image recognition [22] and speech recognition [23]. Some scholars have begun to explore the power of deep learning in capturing media-aware stock movements. For example, Ding, Zhang, and Liu et al. proposed a deep learning method to model both the short- and long-term influences on stock price movements and found that the performance of DNN is better than SVM [24]. Huang et al. applied DNN and CNN to explore the impact of public sentiments in Tweets on stock markets and found that CNN is supreme over DNN because of its network ability to retain spatial information through local connections and weight sharing [25]. Both network structures, however, do not consider the time-series effect of market information. Some researchers began to study the media-aware stock movements using recurrent neural networks (RNNs) [20][26][27]. A good example is the work of Tong et al. which focused on the influence of microblogs on Chinese securities markets by using Long and Short-term Memory (LSTM) model [20]. LSTM is suitable to process the sequential data of uniformly distributed elapsed time. However, market information consists of fundamental and media information. The fundamental data is continuous values sampled with fixed time intervals. In contrast, the media information are discrete values sampled with varying time intervals because of the randomness of news releasing. Such heterogeneity leads to miss valuable information partially or twist the feature spaces. To solve this challenge, we propose a tensor-based event-LSTM (eLSTM). In particular, we model the market information space with tensors instead of concatenated vectors and balance the heterogeneity of two types market information sources with event-driven mechanism.

Table 1. Literatures summary on the influence of news articles toward the stock markets

Category	Literature	Model	Focus		Experiment		
			Information source	Scale	Response	Predictor	Period
Statistical Model and Regression Model	[5]	Statistical Model	Wikipedia	Week	Index	Page view number	12/10/2007-04/30/2012
	[6]	Mutual Information	DJIA	Hour	Price	Message volume	11/12/2012-12/03/2013
	[10]	Linear model	S&P500	Day	Return	Emotion word number	1980-2004
Classical ML-based Model	[11]	KNN	Yahoo	Day	Index	News content	12/06/1997-03/06/1997
	[13]	KNN, SVM	PRNewswire	Minute	Stock trend	News content	04/01/2002-12/31/2002
Deep learning Model	[26]	Neural Network	Reuters, Bloomberg	Week	Return, Volatility, Sharp ratio, Draw-down return	Sentiment	1/2003-2/2014
	[20]	LSTM	Microblogs Chat room	hours	Stock trend	Transaction data and social media content	2015

3. Model Architecture

Stock market is influenced by various information including fundamental data and media information where each of them has irregular data structure. The common strategies in previous studies are to concatenate these heterogeneous data sources into a compound vector. However, the vector-based models treat different information sources as independent features, which lead to the ignorance of the inherent links and fail to capture the stock movements [4]. In addition, the fundamental data are continuous values sampled with fixed time intervals while the media information occurred randomly. Such heterogeneity result in the valuable information being partially missed. Therefore, we represent market information with tensors to preserve the multifaceted and interrelated characters of data. A tensor-based event-driven LSTM (eLSTM) is proposed to capture the nonlinear relation between market information and stock movements. Figure 2 shows the overview of the proposed approach.

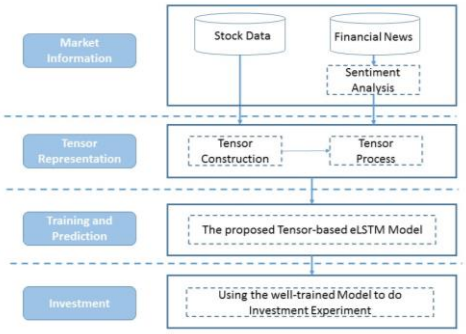


Figure 2. System framework

3.1. Tensor representation

Stock price is a reflection of the expected future cash flows. The release of news information may change investors' expectations and cause stock price to move. Traditional finance believes that the fundamental data reveals the intrinsic value of a company. Modern behavioral finance believes that the release of news affects the expectations of investors which cause market volatility. Stock market information can be roughly categorized into two subgroups, that is, fundamental data and media information. In this study, we extract different information features to construct market space. Specifically,

- Fundamental data: In essence, the stock price reflects a firm's intrinsic value. We have selected eight attributes of a firm to capture its future business value. Each attribute has been shown to have some degree of prediction value in previous studies [7][15][17]. These attributes are, the high price, the low price, the opening price, the closing price, the turnover, the trading volume, the price-to-book and the price-to-earnings ratio.
- Media emotion information: Modern behavioral finance believes that the investors are irrational, and tended to be influenced by the experts' opinions in the media. To capture the media sentiment, we extract the following characteristics: positive media emotion (P_t^+), negative media emotion (P_t^-) and emotional divergence (D_t), denoted by,

$$P_t^+ = \frac{N_t^+}{N_t^+ + N_t^-}, P_t^- = \frac{N_t^-}{N_t^+ + N_t^-}, D_t = \frac{N_t^+ - N_t^-}{N_t^+ + N_t^-} \quad (1)$$

Where N_t^+ (N_t^-) is the number of positive (negative) sentiment words in t^{th}day. In addition, most previous studies capture the media sentiment by utilizing the general emotion word dictionary. However, 73.8% of negative sentiment words in the general dictionary are no longer expressed negative emotional meanings in the financial field [7][28]. For instance, the word bear originally refers to an ursine animal but indicates bad earning returns in finance domain such as bear stock. In this study, we resort to the financial oriented sentiment dictionary[1], which is generously provided by [4] to capture the media emotions.

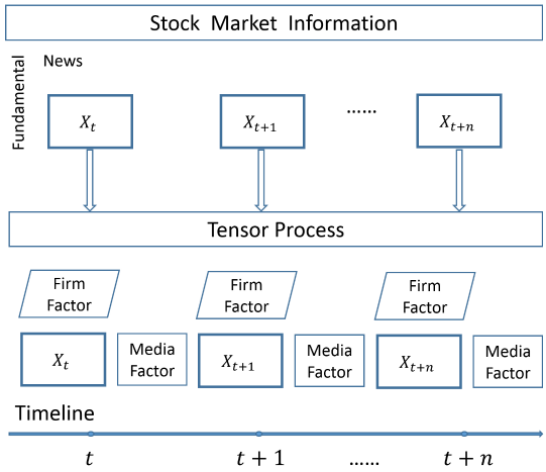


Figure 3. Market information represented with second-order tensor.

After obtaining fundamental and media emotional data, we use a second-order tensor $X_t \in R^{I_1 \times I_2}$ to represent the market information at time *t*. In this way, the

[1] The financial dictionary has been released on the website: http://fife.swufe.edu.cn/BILab/

interconnection of multiple information can be preserved. Figure 3 illustrates an example of a second-order tensor sequence.

The variables I_1 and I_2 are the numbers of fundamental data and media emotional data, respectively. The significance of the element a_{i_1, i_2} in tensor X_t is as follows,

- $a_{i_1,1}$ where $1 \leq i_1 \leq I_1$ denotes the fundamental information feature values.
- a_{2, i_2} where $1 \leq i_2 \leq I_2$ denotes the media emotional information feature values.
- Other elements are set to zeros originally.

Different from the traditional vector-based methods, the market information is represented with a second-order tensor sequence and each subspace represents different information sources. The complementary subspaces are able to capture the correlation of market information. The corresponding stock trends at time t is denoted as $y_t|_{i=t}^{N}$.

3.2 Tensor-based eLSTM model

This study explores the relationship between multiple information sources represented with tensors and the stock movements. In recent research, Long-Short Term Memory (LSTM) performs well when dealing with sequential data [22][23]. In fact, LSTM is a variant of Recurrent Neural Network (RNN) that is capable of handling long-term dependencies with gate structure [29].

In this study, LSTM model is applied to explore the relationship between market information and stock movements. Because media emotion data is with varying time intervals, the standard LSTM utilizes may fail to do the prediction rely on the previous dependencies. For example, it shows that positive news indicates the positive tendency in the previous memory. With the long-time intervals, the model may forget the previous knowledge when processing a similar one. This is because LSTM model is good at dealing with sequential data of uniformly distributed elapsed time like fundamental data. To solve this problem, we propose a novel event mechanism in LSTM to deal with irregular time intervals in the heterogeneous data.

3.2.1 Tensor-based convolution operation

In this study, second-order tensors are utilized to model the complex market information. In the proposed event-driven LSTM, Convolutional LSTM (ConvLSTM) is applied to process tensors for further analyzing instead of fully connected LSTM for 1D vectors [27]. Specifically, with the advantage of local connection and weight sharing, convolution structures of ConvLSTM are able to capture the interactions of different information sources which are modeled as the subspaces of tensors. Essentially, ConvLSTM provides a unique feature to propagate interconnections temporally through each ConvLSTM state. This made it possible for us to process time-series data modeled with tensors.

We define the convolution operation as $'*'$, which extends the vector processing to tensor processing in LSTM, well designed to capture the interrelations inherent market information $X \in R^{I_1 \times I_2}$ for further analysis. The inner structure of convolution operate can be seen in Figure 4.

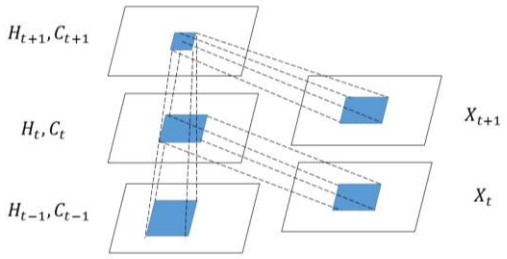


Figure 4. Inner structure of convolution operation

3.2.2 Event-driven LSTM model

Since the input data are represented with tensors, we propose an eLSTM model to solve the invalidated memory problem caused by the non-uniform distributed time intervals. The algorithm details can be seen in Figure 5.

In Figure 5, t represents time period. X_t are the input data in time t. h_{t-1} and C_{t-1} are the output and cell memory obtained in previous time $t-1$. In details, $X_t \in R^{I_1 \times I_2}$ is the market information represented with tensors in time t. The previous output h_{t-1} records the information in hidden layer of previous model, and cell memory C_{t-1} records rules and patterns obtained from the previous market information. In the t period, the forgotten gate f_t controls how much information should be kept or forgotten in the cell memory C_{t-1}. That is f_t can discard the useless information in current time t. Then, by mapping function, temporary memory $\tilde{C}_t$ on market information can be gained from X_t and h_{t-1}. By using input gate i_t, information from the current input data is added to the cell memory C_t. That is, C_t records the useful rules and patterns from information both in previous and current time. The calculations of $\tilde{C}_t$, f_t, i_t and C_t are as follows,

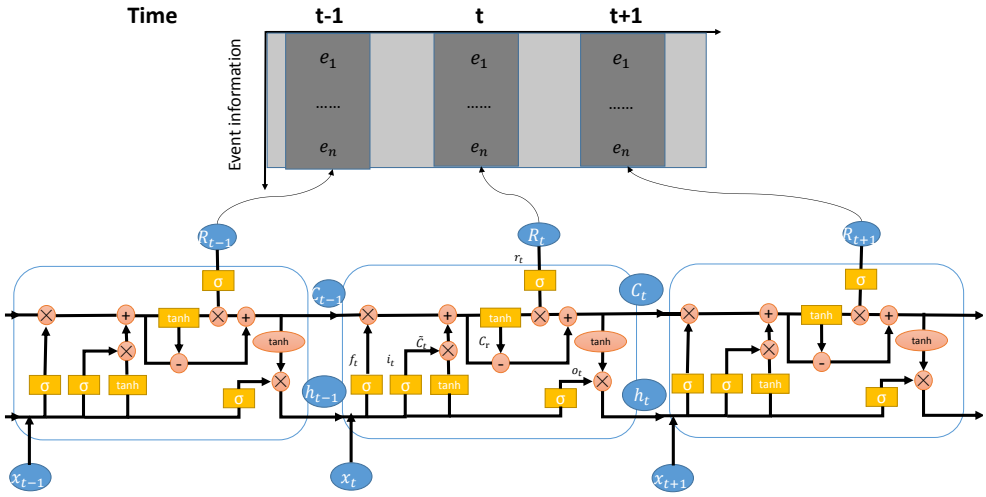


Figure 5. Illustration of the proposed Event-driven Long-Short Term Memory (eLSTM) unit, and its application on analyzing stock market information.

$$\tilde{C}_t = tanh(W_C * X_t + U_C * h_{t-1} + V_c * R_t + b_C) \quad (2)$$

$$f_t = \sigma(W_f * X_t + U_f * h_{t-1} + V_f * R_t + b_f) \quad (3)$$

$$i_t = \sigma(W_i * X_t + U_i * h_{t-1} + V_i * R_t + b_i) \quad (4)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \quad (5)$$

Where $\{W_C, U_C, V_C\}$, $\{W_f, U_f, V_f\}$ and $\{W_i, U_i, V_i\}$ are the parameters of temporary cell, the forgot gate and the input gate of the neural network, respectively. b_C, b_f, b_i are corresponding bias terms. $'*'$ denotes the convolution operator and $'\circ'$ denotes the Hadamard product.

To control the event with irregular time intervals, we use media emotional information R_t to adjust the memory state. With the help of R_t, the model can keep more related memory according to whether an event is in current time t. This part can be seen in Figure 6.

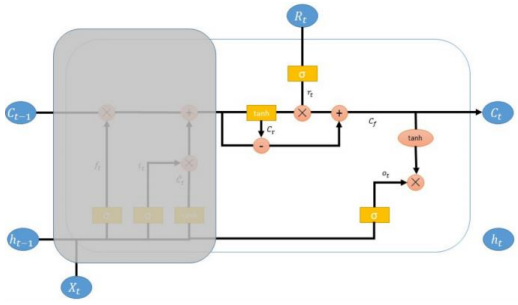


Figure 6. Event-driven details

To achieve this goal, we use non-increasing mapping function, which put the event as the control factor r_t, and use tanh function to get non-relevant events memory C_r from the whole cell memory C_t.

When event happens in current time t, event control factor r_t becomes bigger. More non-relevant information can be discarded and more event related information can be retained. After that, new cell memory C_f is obtained with the help of event-driven mechanism. The calculations of r_t, C_r, C_f are as follows,

$$r_t = \sigma(V_r * R_t + b_r) \quad (6)$$

$$C_r = tanh(C_t) \quad (7)$$

$$C_f = C_t - C_r + C_r \circ r_t \quad (8)$$

Finally, the final cell state C_f and current input data X_t pass through the output gate o_t together and get the output h_t in current time t. The calculations are as follows,

$$o_t = \sigma(W_o * X_t + U_o * h_{t-1} + V_o * R_t + b_o) \quad (9)$$

$$h_t = o_t \circ tanh(C_f) \quad (10)$$

In this study, the stock market volatility is mainly influenced by company fundamental information and media emotional information. To investigate the relationship between market information and stock movements, we use second-order tensor sequence to represent the multi-sources of information. Tensor representation is able to overcome the ignorance of interconnection between two subgroups information with vector-based methods. Additionally, with the continuous fundamental data and discrete media emotion data with non-uniform time intervals, the standard LSTM model finds it is hard to capture the relationships between market information and stock movements. Therefore, we use the eLSTM model to deal with the heterogeneous data and strengthen the impact of news information even with long time intervals.

4. Experimental Evaluation

The purpose of our experiment is to examine the effectiveness of the proposed tensor-based eLSTM model, particularly, the ability to explore the relevance of heterogeneous market information and stock movements. This can improve our understanding of financial markets. Questions of particular interest include the following,

- How to capture the joint effect from multiple information in the stock market? Based on the tensor representation, is it possible to enhance the prediction accuracy by capturing mutual influence and complementary relationship?
- In particular, for the continuous fundamental data and discrete news data with non-uniform time intervals, can eLSTM model strengthen the effectiveness of the event? This may help in proving a new methodology for dealing with heterogeneous data.

We use China Mainland stock market data because market makers are not permitted in Chinese markets. It allows this study to be able to evaluate stock movements in response to multiple market information that is free from the effects from market makers.

4.1 Experimental data

In our experiments, we use stock market data provided by authors in [7], and we extend the data with additional financial news information. In particular, our data consists of fundamentals and financial news.

- Fundamental data: This dataset contains the financial statuses of 100 listed companies in China Securities Index (CSI 100) between January 1, 2015 and December 31, 2015. Since CSI 100 updates the companies in the list every six months, the finalize data consists of 95 companies. The data includes the opening price, the closing price, the high price, the low price, the turnover, the trading volumes, the P/E ratio, and the P/B ratio. According to the research by Li et al., the effectiveness of the historical transaction information on stock market can last 5 days. Therefore, we use 5 days as sliding window in the experiments.
- Media emotion data: The release of important news information affects investors' expectations on the company's future, which bring fluctuations to the stock markets. This dataset contains 12,170 released news related to the 95 companies listed in the CSI 100 during January 1, 2015 to December 31, 2015. The news information gathers from the Chinese financial websites: www.p5h.com. Large amount of comments, clicks and views reflect the importance of the news. After news data has been obtained, we use formula (1) in the section 3.1 to extract the emotional information.

After obtained market information, the tensor representation is applied to construct the input data with a tensor sequence. And then, the data is divided into two sets: a training set and a testing set. The first 9 months of the data were used as the training set and the last 3 months were the testing set. Testing set is used for model evaluation and investment experiments.

4.2 Metrics and tag methods

Followed the previous works in stock prediction [20][24], two criteria, namely the directional accuracy (DA) and matthews correlation coefficient (MCC) are used as the evaluation criteria for the proposed tensor-based eLSTM model. The DA measures the trends prediction compared to the actual changed in the stock prices. MCC is adopted to avoid bias due to data skew. The two criteria are defined as follows,

$$\mathrm{DA} = \frac{n}{N} \tag{11}$$

$$\mathrm{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{12}$$

Where n is the number of correct predictions and N is the total number of predictions, and tp, tn, fp, fn are the number of samples classified as true positive, true negative, false positive and false negative respectively. In particular, there are various methods for 5-days-ahead outcomes [9]. We examine three common methods where the targets are defined in table 2.

Table 2. 5-days-ahead target methods proposed in this paper

Target	Formula
Target 1	$Price_{i+5}^{open} - Price_{i}^{open}$
Target 2	$Price_{i+5}^{close} - Price_{i}^{close}$
Target 3	$Price_{i+5}^{close} - Price_{i}^{open}$

4.3 Effectiveness of the Tensor-based model

When explore the impact of multi-dimensional information of the stock markets, traditional strategy is to concatenate the multiple information into a compound feature vector, where each element is independent from the others, and the interconnection between different dimensions is ignored. Therefore, the vector-based method may lead to the misclassification in trends prediction. In this study, we use tensor representation to preserve the interrelated characteristics of different information dimensions and to find the complementary effect on the stock market. To examine the effectiveness of the tensor-based model, we compare our tensor-based eLSTM model with the vector-based eLSTM model:

• Vector-based eLSTM: We directly feed the concatenated vector into our proposed eLSTM model to do prediction.

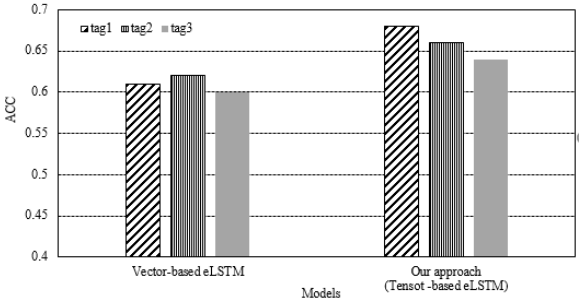


Figure 7. Directional accuracy used to examine the effectiveness of tensor

Figure 7 shows the accuracy of the two models in three tag methods. Compare with the vector-based model, the proposed tensor-based model performs better in all the three target methods. This is because interrelation of multiple data sources that represented with tensor can be retained as much as possible. At the same time, we find that, in accuracy evaluation metric, target 1 and target 2 can better predict the volatility trends of the stocks. This is because investors always collect the information before the opening and after the closing of the market. Therefore, during the market open and the market close, the price can reflect the initial expectations of investors. This finding also supports the point of view that the information in the non-transaction time is able to be reflected by the opening and the closing prices [30].

4.4 Event-Driven effectiveness

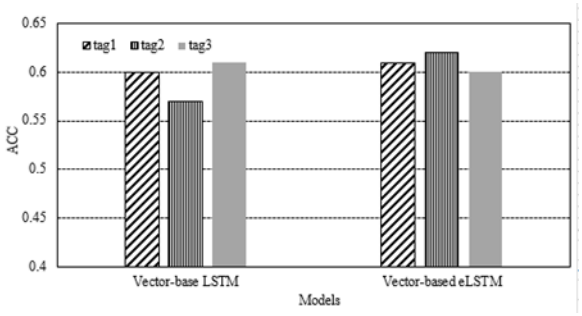
Consider the heterogeneous market information, the dependency from the previous memory may be not significant enough to affect the current output in the standard LSTM. Therefore, this study proposes the eLSTM model to deal with such heterogeneous data. In order to examine the effectiveness of our proposed model, we compare this approach with the following baselines:

• Vector-based LSTM model: We directly feed the concatenated vector into a standard LSTM model to do trends prediction for the stock market.

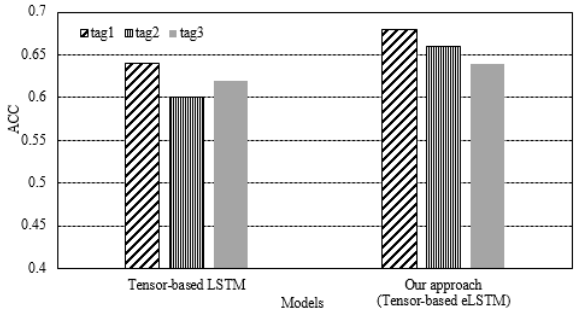
• Tensor-based LSTM model: We use the tensor to represent the multiple market information and then feed the tensor into a LSTM model.

• Vector-based event-LSTM: We directly feed the concatenated vector into our proposed eLSTM model for the prediction.

Comparison of the results on the effectiveness of event-driven models can be seen in Figure 8. The results show that both vector-based models and tensor-based models, the event-driven mechanism in the LSTM can perform better than the standard LSTM. The event-driven method can strengthen the rules and patterns based on the event occurrence in the stock market. That is, in the standard LSTM, the non-uniform time intervals that varies from days to weeks in previous principle may cause the effect of the events get weaker or even lost, and make it insignificant for the prediction. Therefore, eLSTM model is proposed to use whether or not there is an occurrence of the event to capture the related patterns and rules in the sequential data.



（a） Vector-based models



（b） Tensor-based models

Figure 8. Directional accuracy used to examine the effectiveness of Event-Driven

4.5 Tensor-based eLSTM effectiveness

Base on the above examination, it can be found that tensor-based method and event-driven method can effectively capture the effect of multi-sources and heterogeneous data on the stock market. To objectively evaluate the effectiveness of our proposed model, we summarize current models that already show good performances on stock markets and set them as the benchmarks to carry on our experiments:

Table 3. Performance of the models in accuracy and MCC

Model	Target 1		Target 2		Target 3	
	DA	MCC	DA	MCC	DA	MCC
SVM	**0.568**	0.3237	0.518	0.0808	0.526	0.1052
DT	**0.581**	0.4010	0.545	0.1719	0.534	0.1219
ANN	**0.543**	0.2137	0.526	0.0928	0.517	0.0133
LSTM	0.597	0.4318	0.573	0.3733	**0.611**	0.4562
eLSTM	**0.671**	**0.5021**	0.642	0.4843	0.639	0.4732

- Support Vector Machine (SVM): As being one of the effective models for stock forecasting, we select this model to handle our data.
- Decision Tree (DT)：In stock market prediction study, this model has been proved to outperform the others. Therefore, we use it as one of our benchmarks.
- Artificial Neural Network (ANN): ANNs can perform well in classification, regression and pattern recognition problems. Therefore, many studies used this model to do stock prediction [12]. We also select it as a baseline.
- LSTM: LSTM has excellent performance in processing time-series data. Here, we use a standard LSTM to predict stock trends as the baseline.

The results of the experiment are shown in Table 3. Comparing with SVM, DT and ANN models, LSTMs show a better performance in all 3 tags. This can be seen that the Long and Short time Memories algorithm is able to capture the rules and patterns in the sequential data. Therefore, the LSTM model is suitable for the prediction of stock market volatility. Furthermore, when comparing with LSTM, eLSTM shows a better performance. This indicates that the proposed model can capture the joint effects of multiple market information and strengthen the impact of events for better prediction.

4.6 Investment

We design and implement a tensor-based stock market information analyzer based on the proposed eLSTM framework. In this section, we compare the performance of our Top-N (N = 5, 10 ,15 and 20 means the number of highest-performance stocks) analyzer with three state-of-the-art trading algorithms: AZFinText [8], eMAQT [8] and TeSIA [4].

Similar to the baseline algorithms, we set the investment budget at RMB 10,000 and compare the daily earning of these approaches over three months. The incomes from the daily investment are shown in Figure 9. While CSI 100 index decreased by 5.3% (from 3176 to 3007), eMAQT return reached 121.11%. TeSIA achieved the best return 150.27% among the baselines, while the performance of proposed top-5 strategy algorithm in this study yielded a remarkable return of 189.10% over the three months, which is better than the state-of-the-art algorithms.

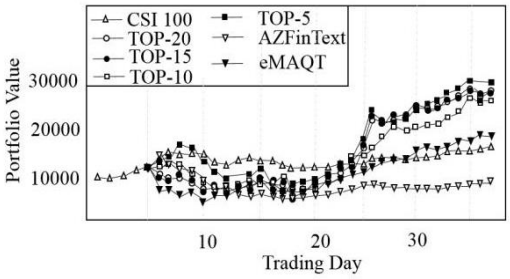


Figure 9. Investment comparison

5. Conclusion and Future Work

In this article, we propose a tensor-based eLSTM model to capture the effects of multi-dimensional data on stock market movements. Compare with traditional vector-based models: the SVM, the Decision Tree and the ANN, we find that tensor representation can identify multiple information factors and capture the intrinsic relations in the stock market. In addition, LSTM model does outperformance the others because of its ability to deal with sequential data. However, as the input data includes both continuous fundamental data and news data with non-uniform time intervals, the standard LSTM model may fail to capture the effectiveness of the heterogeneous data due to irregular elapsed times. To solve this, we introduce an event-driven algorithm to take the irregular time into account. Even with a long interval, the algorithm can call for special patterns according to the occurrence of the events. However, in this study, we get the emotional tendency of financial news by using simple dictionary-based method [10]. This approach ignores the impact of transitional words, comparative words, and negative words such as ‘but’ and ‘not’ and does not consider the syntax analysis. In the future work, we would find some efficient way to quantify text information and capture the context structural relationship.

More important, this study provides a new methodology for dealing with heterogeneous data structure with both continuous data and non-uniform distributed

elapsed time intervals data. This tensor-based eLSTM framework can be applied to many other situations. For example, the continuous growth data of daily crop with discrete rainfall data in agriculture, and the continuous daily index data with uncertain option transaction data in finance field. We would be very interested in studying the effectiveness of this method in these fields.

6. Acknowledgments

This work has been supported by grants awarded to Dr. Qing Li from the National Natural Science Foundation of China (NSFC) (71671141, 71401139, 60803106, 61170133), Fundamental Research Funds for the Central Universities (JBK1707081, JBK 171113, JBK 170505, JBK 151128, JBK 18FG08) and the Sichuan Science and Technology Innovation team training plan (2011JTD0028). It also has been partially funded by the Key Laboratory of Financial Intelligence and Financial Engineering of Sichuan Province (FIFE).

7. References

[1] J.B.D. Long, A. Shleifer, L.H. Summers, and R.J. Waldmann, "Noise trader risk in financial markets", J. Political Ecom., vol. 98, no. 4, (1990), pp. 703-738.
[2] E.F. Fama, "The behavior of stock-market prices", J. Bus., vol. 38, no. 1, (1965), pp. 34–105.
[3] M. Rechenthin and W.N. Street, "Using conditional probability to identify trends in intra-day high-frequency equity pricing", Physica A Stat. Mech. Appl., vol. 392, no. 24, (2013), pp. 6169-6188.
[4] Q. Li, Y. Chen, L. L. Jiang, P. Li, and H. Chen, "A tensor-based information framework for predicting the stock market," ACM Trans. Inf. Syst., vol. 34, no. 2, (2016), pp. 11:1-11:30
[5] I. Zheludev, R. Smith, and T. Aste, "When can social media lead financial markets?", Sci. Rep., vol. 4, (2014), pp. 1-12.
[6] H.S. Moat, et al., "Quantifying Wikipedia usage patterns before stock market moves", Sci. Rep., (2013), pp. 1-15.
[7] 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, (2014), pp. 826–840.
[8] R.P. Schumaker and H. Chen, "A quantitative stock prediction system based on financial news", Inf. Process. Manag., vol. 45, no. 5, (2009), pp. 571–583.
[9] B. Weng, M.A. Ahmed, and F.M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources", Expert Syst. Appl., vol. 79, (2017), pp. 153-163.
[10] P.C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: quantifying language to measure firms' fundamentals", J. Finance., vol. 63, no. 3, (2008), pp. 1437–1467
[11] B. Wang, H. Huang, and X. Wang, "A novel text mining approach to financial time series forecasting," Neurocomputing., vol. 83, (2011), pp. 136–145.
[12] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2 (2011), pp. 1–8.
[13] M.A. Mittermayer, and G. F. Knolmayer, "Newscats: A news categorization and trading system", in Proc. IEEE 6th Int. Conf. Data Mining., (2006), pp. 1002-1007.
[14] T. Thanh, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Tensor representation in high-frequency financial data for price change prediction", arXiv preprint arXiv:1709.01268, (2017).
[15] R.A. Haugen, and N.L. Baker, "Commonality in the determinants of expected stock returns", J. Financ. Ecom., vol. 41, no. 3, (1996), pp. 401-439.
[16] A. Shleifer and R.W. Vishny, "The limits of arbitrage", J. Finance., vol. 52, no. 1, (1997), pp. 35–55.
[17] E.F. Fama and K.R. French, "Common risk factors in the returns on stocks and bonds", J. Financ. Ecom., vol. 33, no. 1, (1993), pp. 3–56.
[18] M.Z. Frank, and W. Antweiler, "Is all that talk just noise? The information content of internet stock message boards", J. Finance., vol. 59, no. 3, (2004), pp. 1259-1294.
[19] X. Luo, J. Zhang, W. Duan, "Social media and firm equity value", Info. Syst. Research., vol. 24, no. 1, (2013), pp. 146-163.
[20] T. Sun, et al., "Predicting Stock Market Price Returns using Microblogs sentiment for Chinese Stock Market", ICBDC, (2017), pp. 87-96.
[21] G.E Hinton, and R.R Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science., vol. 313, no. 5786, (2006), pp. 504-507.
[22] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Int. Conf. Neural Inform. Process. Syst., (2012), pp. 1097–1105.
[23] G.E Hinton, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,", IEEE Signal Process. Mag., vol. 29, no. 6, (2012), pp. 82–97.
[24] X. Ding, Y. Zhang, T. Liu, J. Duan. "Deep Learning for Event-Driven Stock Prediction", IJCAI, (2015), pp. 2327-2333.
[25] Y. Huang, K. Huang, Y. Wang, H. Zhang, J. Guan, and S. Zhou, "Exploiting Twitter Moods to Boost Financial Trend Prediction Based on Deep Network Models", ICIC., (2016), pp. 449-460.
[26] Q. Song., A. Liu, and S.Y. Yang, Neurocomputing., Available: http://dx.doi.org /10.1016/j. neucom, (2017).
[27] X.J. Shi, et al., "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", INIPS, (2015).
[28] Q. Li, J. Wang, F. Wang, P. Li, L. Liu and Y. Chen, "The role of social sentiment in stock markets: a view from joint effects of multiple information sources", Multimedia Tools and Applications., vol. 76, no. 10, (2017), pp. 1-31.
[29] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neurocomputation., vol. 9, no. 8, (1997), pp. 1735–1780.
[30] K. Tissaoui, "The Intraday Pattern of Trading Activity, Return Volatility and Liquidity: Evidence from the Emerging Tunisian Stock Exchange", J. Ecom and Financ., vol. 4, no. 5, (2012), pp.156-176.