SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2021, Vol. 64 192102:1–192102:13
https://doi.org/10.1007/s11432-020-3064-4

Quantifying the effects of long-term news on stock markets on the basis of the multikernel Hawkes process

Xiao DING, Jihao SHI, Junwen DUAN, Bing QIN & Ting LIU*

Research Center for Social Computing and Information Retrieval, Harbin 150001*, China*

Received 12 March 2020/Accepted 3 June 2020/Published online 27 July 2021

Abstract Recent studies have revealed that long-term financial news can affect on stock markets. However, previous research mainly focuses on modeling the short-term effects of financial news and suffers from the weak ability of quantifying the time-decaying influence of financial news. To fill this gap, this study introduces the Hawkes process to estimate the time-decaying influence of long-term financial news. However, the performance of the conventional Hawkes process is sensitive to the choice of kernel functions. Hence, we propose a novel multikernel-powered Hawkes process framework, which uses multiple kernels to model different time-decaying rates, thus alleviating the instability of our proposed Hawkes process based prediction model. Experimental results show that the proposed framework yields state-of-the-art stock market prediction accuracies on 515 listed companies and gains more profits in market trading simulation compared with baseline methods. News-based stock prediction can complement studies on price-volume-based stock prediction.

Keywords Hawkes process, multikernel function, stock market prediction, long-term news effects

Citation Ding X, Shi J H, Duan J W, et al. Quantifying the effects of long-term news on stock markets on the basis of the multikernel Hawkes process. Sci China Inf Sci, 2021, 64(9): 192102, https://doi.org/10.1007/s11432-020-3064-4

1 Introduction

Financial news can help people understand the volatility of stock markets [1–4]. As shown in Figure 1, some breaking news can quickly affect stock prices in a short period. Meanwhile, their long-term influence may remain and result in consequent relevant news events. For example, on January 14th, Reuters reported that Samsung offered to buy BlackBerry for as much as $7.5 billion, thus stimulating the stock price of BlackBerry to soar on that day. However, in the following days, both companies denied that they were having talks of any possible takeover; thus, the influence of the good news decayed in time, and the stock price continued to fall. Although the acquisition event was later denied, it still had a certain positive effect on the subsequent stock price of BlackBerry. Previous research focuses on short-term (one day) effects of financial news [5]. However, the consideration of time-decaying influence on long-term financial news still leaves a substantial gap to explore.

To fill this gap, Ding et al. [6] proposed a novel CNN-based framework to model the short- and long-term effects of news events jointly on stock price movement. Hu et al. [2] and Xu et al. [7] maximized the temporal attention mechanism to fix long-term series dependencies. However, these studies were limited by the inability to quantify the time-decaying influence of long-term financial news. The quantification of time decay can provide a fine-grained investigation on the effects of news on stock markets and can serve as a reasonable explanation for the prediction results.

Hence, we investigate the Hawkes process [8] to estimate the effects of long-term financial news. The Hawkes process is a self-exciting temporal point process with three intrinsic natures, which can facilitate our task. (1) Self-exciting mechanism. Each past event contributes to the arrival rate of subsequent events

* Corresponding author (email: tliu@ir.hit.edu.cn)

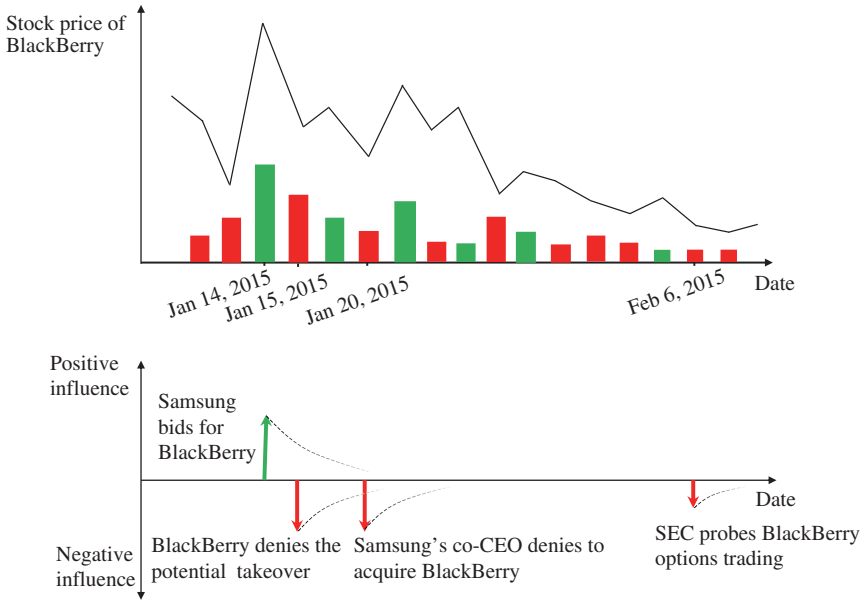

Figure 1 (Color online) The influence of historical news decays with the increase of time. On January 14th, Reuters reported that Samsung had offered to buy BlackBerry for as much as $7.5 billion, which stimulated the stock price of BlackBerry to soar on that day. But in following days, both companies denied they were in talks with respect to any possible takeover, thus the influence of the good news was time-decaying and the stock price continued to fall. We notice that although the acquisition event was later denied, it will still have a certain positive effect on the subsequent stock price of BlackBerry.

in the future. In other words, the arrival rate of the next event depends on past events. In our task, the stock movement prediction of individual firms depends on the historical financial news events during a period. (2) Accumulation mechanism. The Hawkes process makes predictions through the accumulation of historical events. This feature enables our proposed model to learn the cumulative effects of long-term financial news. The accumulation of long-term financial news can consider informative news more than short-term based methods. (3) Quantifying the time-decaying influence. The kernel function in the Hawkes process enables our proposed model to quantify the time-decaying influence of long-term news with the increase in time.

First, we follow Duan et al. [9] to learn firm-specific representations of financial news for each individual firm. Then, by gathering these news representations within the same day, we can obtain daily news representations. Afterward, we use the kernel function in the Hawkes process to quantify the time-decaying influence of financial news. As pointed out by Lima et al. [10], kernel choice is critical to the power of the Hawkes process because different kernel function choices can lead to different time-decaying rates. To alleviate this issue, we devise an optimal multikernel selection procedure to enhance the robustness of our prediction model. Lastly, we combine the daily vector representations of historical financial news with corresponding quantified time-decaying influence to predict the volatility of the stock market.

The main contributions of this study are two-fold.

• We propose to quantify the time-decaying influence of financial news on the stock market on the basis of the Hawkes process.

• We explore multiple kernels for the Hawkes process, substantially enhancing the robustness of our model in comparison with single-kernel methods.

Experimental results on the stock market show that our approach can effectively model the time-decaying influence of long-term financial news and achieve better prediction performances compared with state-of-the-art baseline methods.

2 Background

2.1 Problem definition

Cumulative abnormal return. In this study, we focus on the task of cumulative abnormal return

prediction, which studies the effect of news information toward a specific firm. Formally, the actual return R_{t_n} of a firm on the trading day t_n is calculated by

$$R_{t_n} = \frac{\text{close}_{t_n} - \text{close}_{t_{n-1}}}{\text{close}_{t_{n-1}}}, \tag{1}$$

where close_{t_n} is the closing price of the trading day t_n, and $\text{close}_{t_{n-1}}$ is the closing price of the former trading day t_{n-1}. Abnormal return AR_{t_n} is the difference between the actual return R_{t_n} of a stock and its expected return $\hat{R}_{t_n}$, which is given as

$$\text{AR}_{t_n} = R_{t_n} - \hat{R}_{t_n}, \tag{2}$$

where the expected return $\hat{R}_{t_n}$ can be approximated by the stock index, such as S&P 500 index. The cumulative abnormal return is calculated by cumulating the daily abnormal return in a fixed window. Following Duan et al. [9], we use a three-day window (−1, 0, 1), which is denoted as CAR_3, with the trading day t_n at the center of the three-day window. Formally,

$$\text{CAR}_3 = \text{AR}_{t_{n-1}} + AR_{t_n} + \text{AR}_{t_{n+1}}. \tag{3}$$

Prediction. Given a trading day t_n, the task is to model the influence of its historical financial news sequences for predicting the cumulative abnormal return of that day. We define long-term historical financial news as news over the past month, and short-term news as news on the past day of the trading day t_n (the same setting as [5,6]). To simplify, we take the length of the long-term historical news window that equals to 30 as an example $[t_{n-30}, \ldots, t_{n-2}, t_{n-1}]$. Formally, by using the target-specific document representation method [9], we can acquire the daily news representations $[r_{t_{n-30}}, \ldots, r_{t_{n-2}}, r_{t_{n-1}}]$. Then, we combine them with corresponding weights (i.e., time-decaying rates) calculated by the multikernel function in the Hawkes process to obtain the final representations of historical news d_f^c about the company c. The probability that d_f^c has positive or negative effects on the stock prices of company c is predicated via logistic softmax regression,

$$p(y|d_f^c) = \text{softmax}\left(W \cdot d_f^c + b\right), \quad y \in \{0, 1\}, \tag{4}$$

where W is a weight matrix and b is a bias vector. $p(y = 0|d_f^c)$ and $p(y = 1|d_f^c)$ indicate the probability of CAR_3 being negative and positive, respectively.

2.2 Hawkes process

The Hawkes process is a self-exciting point process, which has been applied in many fields, such as seismology literature [11], popularity dynamics [12], and invasive species management [13]. The simplest form of point process is the homogenous Poisson process, which assumes that events are independent from each other, and the arrival rate of events is static. Considering the influence of historical events, the Hawkes process models the arrival rate of new events. The intensity function λ_k is considered the number of events in an infinitesimal interval and is also known as the arrival rate of new events. Its formulation is as follows:

$$d_f^c = \mu_{t_{n-1}} + \sum_{i=2}^{30} \mu_{t_{n-i}} \phi\left(t_{n-1} - t_{n-i}\right). \tag{5}$$

In our task settings, d_f^c is the final representation of the historical news sequence $[t_{n-30}, \ldots, t_{n-2}, t_{n-1}]$. $\mu_{t_{n-1}}$ is a vector which stands for the representation of the news released on the date t_{n-1} closest to the trading day t_n. $\mu_{t_{n-i}}$ is the historical news representation which implies the positive or negative effect towards the stock market. t_{n-i} represents the released date of the historical news. The function $\phi(\cdot)$ is a kernel function to model the decaying influence.

The Hawkes process captures three key factors: Self-exciting mechanism — Each past event contributes to the arrival rate of subsequent events in the future. In other words, the arrival rate of the next event depends on past events. Ding et al. [5] showed that news that happened on t_{n-1} has a remarkable influence towards the trading day t_n. Accumulation mechanism — By accumulating long-term historical news, we consider more historical information in comparison with short-term based methods. Time decay effect — The influence of historical news decays as time goes.

Inspired by Cao et al. [14], we construct an end-to-end supervised deep learning framework. We use the prediction result to supervise the entire process, which is different from those unsupervised generative process methods [12, 15] by utilizing the Hawkes process.

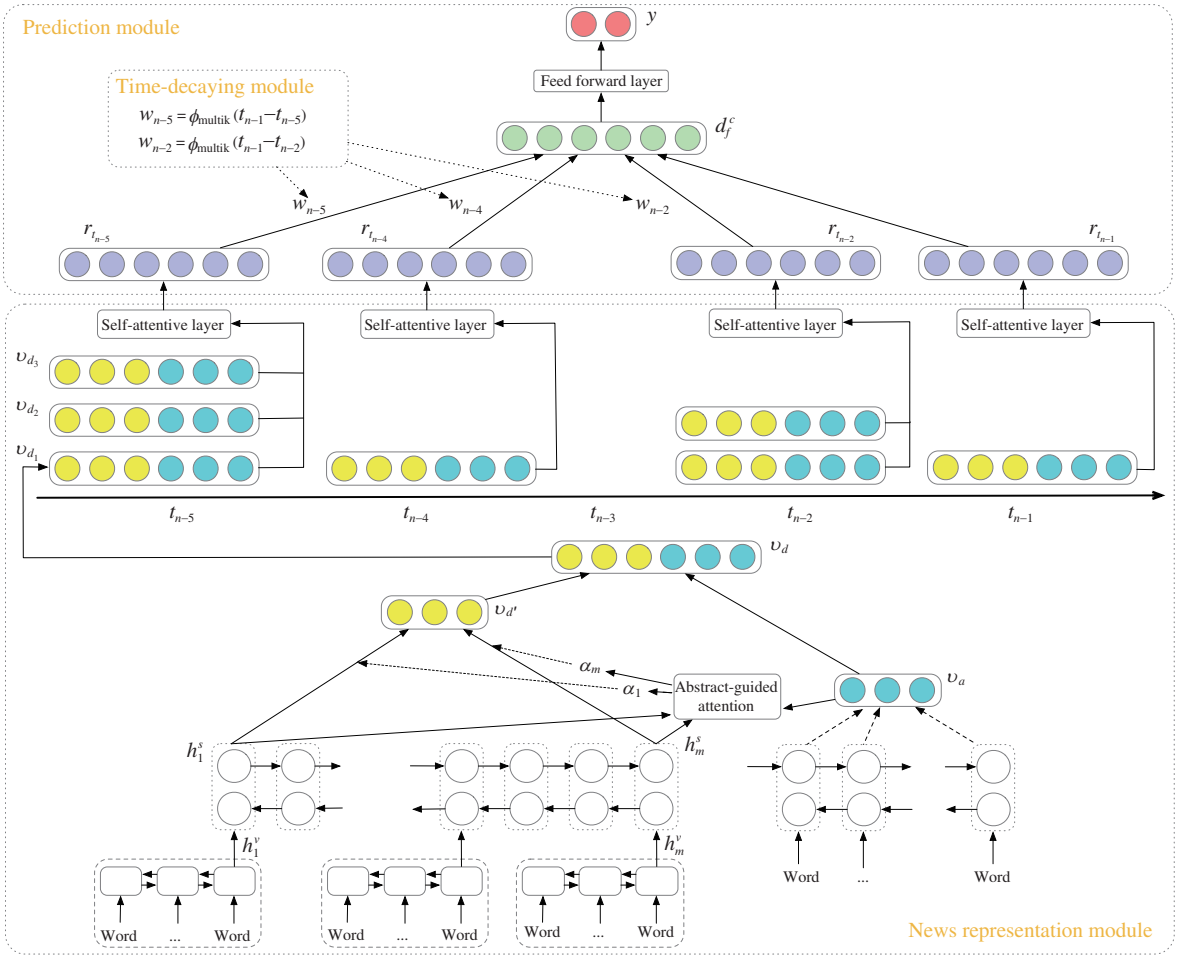

Figure 2 (Color online) Framework of stock prediction in combination with the Hawkes process. It omits part of the lines in case of chaos. t_{n-5} has three financial news pieces, whereas t_{n-3} has no news about specific firm c.

3 Multikernel Hawkes process model

As shown in Figure 2, our framework comprises three components. First, we learn the daily news representations using a target-specific, abstract-guided method. Then, we calculate the decaying factors of historical news using the multikernel function. Subsequently, we combine the news representation and decaying factors to acquire the final representation of the sequences of historical news. Lastly, we use a softmax layer to predict the stock movement of the firm.

Given a trading day, the input is a series of historical news which happens in a fixed window. If no news is available on a specific day, we take null in that position. For example, if the window length is 5, then the consecutive time span is 5 days. During the time span, we ensure that at least one day is filled with news. The output is the cumulative abnormal return of the trading day.

3.1 News representation

Sometimes a news piece may mention more than one firm, and its effects on different firms can vary [9,16]. Therefore, we must learn distinct representations of the same news piece for different firms. In addition, we believe that the news body contains more informative background knowledge compared with the news title; thus, the news body is beneficial for stock prediction. Hence, we follow Duan et al. [9]'s approach of using target-specific, abstract-guided document representation to learn distributed vectors for financial news documents. To reduce document noise, we learn a target-specific representation of the abstract to guide us in selecting the most informative sentences for the entire document.

First, we learn a target-specific representation v_a for the abstract of a news piece via conditional encoding [17]. We use bi-directional long short-term memory (Bi-LSTM) [18] structure as the basic

representation learning model because the news abstract is a sentence. Among various neural networks for encoding sentences, Bi-LSTM has been a dominant method, providing state-of-the-art results in language modelling [19] and syntactic parsing [20]. To encode specific information into v_a, we use an embedding vector of the individual firm as the initial state vector for the sentence-level Bi-LSTM. The vector for the individual firm is initialized by averaging the words of its constituents and fine-tuned during training.

Then, we encode each sentence (the abstracts are not considered) in the document to a context-sensitive representation h_i^s using a hierarchical structure [21]. A sentence-level Bi-LSTM is first used to encode words into a hidden state vector h_i^v. In the document level, given the sentence embedding $\{h_1^v, h_2^v, \ldots, h_m^v\}$ for sentences $\{s_1, s_2, \ldots, s_m\}$ in a document, respectively, the same Bi-LSTM structure (with a different set of model parameters) is used to obtain hidden states $\{\overrightarrow{h_1^s}, \overrightarrow{h_2^s}, \ldots, \overrightarrow{h_m^s}\}$ and $\{\overleftarrow{h_1^s}, \overleftarrow{h_2^s}, \ldots, \overleftarrow{h_m^s}\}$, respectively. For each sentence s_i, the forward and backward hidden vectors are averaged to provide a single hidden state embedding h_i^s, which contains the semantic composition of words and its context information.

Afterward, we use the target-specific abstract vector v_a to guide the calculation of attention weights α_i for each sentence h_i^s in the document. By using the weights, the model can not only address the challenge of informative sentence selection, but also provide an evidence about the prediction results. The weight score u_i^s shows how much attention should be placed on sentence h_i^s:

$$u_i^s = v^{\mathrm{T}} \tanh\left(W_a \left[v_a : h_i^s\right] + b\right), \tag{6}$$

where v is a vector which projects the result to a scalar; W_a is a weight matrix; b is a bias vector and [:] denotes the concatenating operation.

The normalized weight score α_i and the abstract-guided document representation $v_{d'}$ are computed from (7) and (8), respectively. The final document representation is $v_d = [v_a : v_{d'}]$.

$$\alpha_i = \frac{\exp\left(u_i^s\right)}{\sum_i \exp\left(u_i^s\right)}, \tag{7}$$

$$v_{d'} = \sum_i \alpha_i h_i^s, \tag{8}$$

where u_i^s is a weight score and h_i^s is a hidden state embedding as mentioned above.

More than one news can be obtained in a day; thus, the model adopts a self-attentive neural network [22] to assign different weights (Eq.(9)) to news documents $\{d_1, d_2, \ldots, d_n\}$ with respect to our prediction goal.

$$u_i^d = v^{\mathrm{T}} \tanh\left(W^{(m)} v_{d_i} + b^{(m)}\right), \tag{9}$$

where v is a vector which projects the result to a scalar; v_{d_i} is the representation of document d_i; $W^{(m)}$ is a weight matrix and $b^{(m)}$ is a bias vector. The normalization process is the same with (7) but with different parameters. Lastly, we acquire the representation of daily news r_t on the date t.

3.2 Learning time-decaying influence with the multikernel function

Ding et al. [6] showed diminishing effects of historical financial news on stock market volatility. To quantify such effects, we propose using the Hawkes process to model the time-decaying influence through a parametric kernel function. The parametric kernel function $\phi(t)$ determines the rate of the time-decaying influence. Therefore, learning a proper kernel function is essential for our task.

However, the selection of an appropriate kernel function is intractable, and thus commonly dependent on the prior domain knowledge. For example, in studying causal influences in a blogosphere [23], the decaying influence is at a quick rate, which can be formulated by an exponential kernel function $\phi_{\exp}(t) = \mathrm{e}^{-\theta t}$. By contrast, in aftershock prediction [11], the decaying influence has a slow rate, which can be formulated by a power-law kernel function $\phi_{\mathrm{pow}}(t) = (t + c)^{-(1+\theta)}$. When we have minimal advance knowledge, the Gaussian kernel is an expressive means to encode prior knowledge [24], and a sequence of complex real-valued variables can be explained by a compositional kernel with base kernels [25]. The mathematic formalization of the Gaussian kernel function is shown as

$$G(t) = \phi_{\mathrm{Gaussian}}(t) = \mathrm{e}^{-t^2/2\sigma^2}, \tag{10}$$

where t is a delta interval between release dates, σ is a bandwidth, and $G(t)$ and $\phi_{\text{Gaussian}}(t)$ are two different expressions of the same formula. We take the Gaussian kernel as a base kernel, and then a compositional kernel (i.e., Gaussian multikernel) comprises five Gaussian kernels with different bandwidths to model different time-decaying rates. Furthermore, we incorporate the exponential kernel and power-law kernel into the Gaussian multiple kernel function as

$$\phi_{\text{multik}}(t) = \beta_{\text{exp}}\phi_{\text{exp}}(t) + \beta_{\text{pow}}\phi_{\text{pow}}(t) + \sum_{u=1}^{5} \beta_u G_u(t), \tag{11}$$

where t is a delta interval between release dates; $G_u(t)$ is a Gaussian kernel $G(t)$ with a specified bandwidth σ; and β_{exp}, β_{pow}, and β_u are positive parameters which satisfy $\beta_{\text{exp}} + \beta_{\text{pow}} + \sum_{u=1}^{5} \beta_u = 1$. β_{exp}, β_{pow}, and β_u are first given random initial values and then are updated in the training process.

3.3 Prediction via the Hawkes process

We have obtained daily news representations and learned the time-decaying influence of financial news from Subsections 3.1 and 3.2. Each daily news can contribute to the stock movements on the trading day t_n, and the news reported on the date t_{n-1} can be viewed as the strongest indicator [6]. With the incorporation of time-decaying influence, the representation d_f^c of the news sequence is assembled by a weighted sum pooling mechanism:

$$d_f^c = r_{n-1} + \sum_{h:t_h<t_{n-1}} r_h w_h, \tag{12}$$

$$w_h = \phi_{\text{multik}}\left(t_{n-1} - t_h\right), \tag{13}$$

where r_h is the vector representation of news reported on date t_h; t_n is the trading day on which we want to predict the market movements; w_h is the weight calculated by the multiple kernel function to quantify the time-decaying influence; d_f^c is the representation of financial news sequence.

Lastly, taking the representation d_f^c as input, we use a single feed-forward layer to predict the trend of cumulative abnormal return of individual firm c.

$$p(y|d_f^c) = \text{softmax}\left(W \cdot d_f^c + b\right), \quad y \in \{0, 1\}, \tag{14}$$

where W is a weight matrix and b is a bias vector.

3.4 Training details

The model is trained in a supervised manner by minimizing the cross-entropy of the prediction result, whose loss function is given as follows:

$$\min - \sum^{T} [y \cdot \log(p) + (1 - y) \cdot \log(1 - p)], \tag{15}$$

where y is 1 when the label of training data is rising, otherwise y is 0. p is the probability that the predicted label is rising, and T is the number of training instances. We use Adam [26] to update the entire set of parameters by taking the derivative of the loss through backpropagation.

The word vectors are initialized from pretrained word embeddings, which are trained on Bloomberg and Reuters corpus [6] and fine-tuned during our model training. We choose the parameters that can achieve the best micro-F1 result on the development set for the final test. Early stopping is also applied to avoid overfitting issues and save training time. We empirically set the embedding length as 100, the length of hidden layer as 100 and the learning rate of Adam as 0.0005 for all baseline models and our model. The hyperparameter σ^2 in the Gaussian function obtains values in [0.005, 0.05, 0.5, 5, 50], and we then investigate the influence of different number of Gaussian functions in Subsection 4.6.

Table 1 Statistics of datasets

	Training	Development	Test
# records	18042	996	2003
Time interval	22/10/2006–14/04/2014	15/04/2014–01/09/2014	02/09/2014–26/08/2015

4 Experiments

4.1 Data

We use publicly available financial news from Reuters between 22/10/2006 and 26/08/2015 [6]. We also obtain the AMEX, NYSE and NASDAQ composite indices from Yahoo, as well as the stock prices of firms. We calculate the expected return $\hat{R}_t$ using equally-weighted market indices. As shown in Table 1, we split the dataset into training, development, and test sets; such splitting process is the same as that in Duan et al. [9].

4.2 Baselines and evaluation metrics

We compare our approach with three categories of baseline methods, with accuracy (%) as the evaluation metric. We first compare the performance of long- and short-term-based prediction methods. Then, we compare the performance of different long-term-based prediction methods (model-CNN, model-no-weight and model-temporal-attention). Lastly, we compare the performance of single-kernel (model-exp, model-pow and model-Gaussian) and multikernel Hawkes process methods.

short-term. Duan et al. [9] presented the current state-of-the-art accuracies for CAR prediction by using target-specific, abstract-guided, document-level representations of intraday financial news for stock movement prediction.

model-no-weight. This baseline uses the same prediction framework as our proposed approach, except that the final long-term news sequence representation is calculated by the summation of equally-weighted representations of daily news.

model-CNN. Ding et al. [6] split the historical temporal span into three parts: a day, a week and a month. Then they used a deep convolutional neural network to model the combined influence of long-term and short-term events on stock price movements.

model-temporal-attention. Hu et al. [2] and Xu et al. [7] adopted the temporal-level attention to distinguish the influence of financial news reported on different days. Both pay attention on the news content without considering the timeliness of news. This baseline uses an attention mechanism to learn different weights for each long-term news on the basis of the frameworks proposed by Hu et al. [2] and Xu et al. [7].

model-exp, model-pow, multik (Gaussian). These kernel based baseline methods can model the time-decaying influence of financial news in different rates. The exponential kernel function (model-exp) can model the time-decaying influence at a quick rate, whereas the power-law kernel function (model-pow) models the influence at a slow rate. We also use a composition of multiple Gaussian kernels (multik (Gaussian)) as a baseline method.

4.3 Overall results

The experimental accuracy results of all baselines and our approach are shown in Table 2. All baselines and our model use the same dataset and settings mentioned in Subsection 4.1. We conduct an analysis by comparing these baseline methods with our approach in detail.

(1) Comparing Duan et al. [9] with our approach, the short-term-based method performs poorly because in our approach, we consider more news information. Despite the relatively weaker effects of long-term news, the volatility of the stock market is still affected by them. Additional information can help the system make more accurate predictions.

(2) Compared with different long-term based prediction methods (model-no-weight, model-CNN, and model-temporal-attention), our approach achieves the best performance. Model-no-weight can use long-term news information and thus achieves better performance compared with short-term-based baseline methods. However, giving equal weights to each long-term news is unreasonable. Although model-CNN [6] can model the combination effects of long-, middle- and short-term news for stock market, it cannot provide an explicitly quantitative analysis of such effects because it cannot provide a concrete score

Table 2 Experimental accuracy results of stock movement prediction[a)]

Method	Accuracy (%)
short-term (Duan et al. [9])	52.32
model-no-weight	53.25
model-CNN (Ding et al. [6])	53.72
model-temporal-attention (Hu et al. [2])	54.12
model-temporal-attention (Xu et al. [7])	54.57
model-exp (ours)	55.17
model-pow (ours)	55.32
multik (exp+pow) (ours)	55.97
multik (Gaussian) (ours)	56.02
multik (exp+pow+Gaussian) (ours)	**56.12**

a) The best result is in bold.

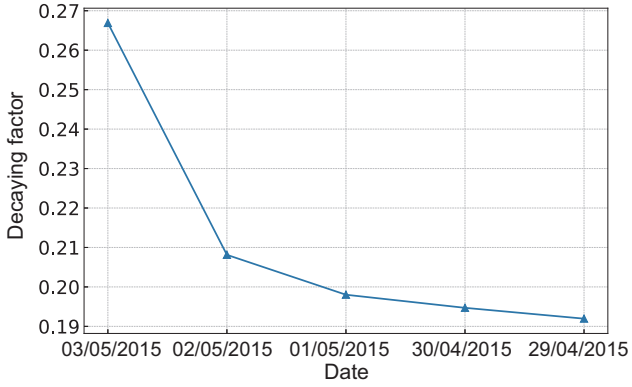

Figure 3 (Color online) Quantification of time-decaying influence on historical consecutive days in an interval of 6 days (29/04/2015–04/05/2015) using multikernel Hawkes process.

to quantify the influence of each news. model-temporal-attention [2,7] learns different effects of long-term news through an attention mechanism but only considers the news content and ignores the decline of the news influence caused by the timeliness. Our proposed approach quantifies the time-decaying influence of news by integrating the content and timeliness of the news.

(3) The multikernel Hawkes process based method outperforms all single-kernel Hawkes process-based methods, thus demonstrating the power of the optimal multikernel-based operation. Each kernel function has its own specific advantage, whereas multiple kernels can gather the advantages of each kernel function.

4.4 Case study

To analyze the effectiveness of our proposed approach, we draw a curve that shows the quantification of time-decaying influence in Figure 3. We want to predict the cumulative abormal return of a specific firm on 05/05/2015. In our proposed model, the weight on the closest date 04/05/2015 is 1. The figure shows different weights of historical news in an interval of 6 days (here, we omit the closest date 04/05/2015 in the figure). During the time interval, 02/05/2015 is Saturday and 03/05/2015 is Sunday. Previous research considered these two days equally weighted toward prediction and merged them together as a whole, similar to news occurring on the same date [5]. However, different decaying factors can be observed in these two days. Moreover, the effects of news events drop considerably the day after they happen, but their influences can still continue for some time. These long-term effects are also useful for predicting the stock movements but are often overlooked.

4.5 Accuracy vs. different time intervals

We also investigate the performance of long-term news-based methods during a period at different time intervals (from 2 days to 30 days). The accuracy time interval curves are shown in Figure 4. We can find the following.

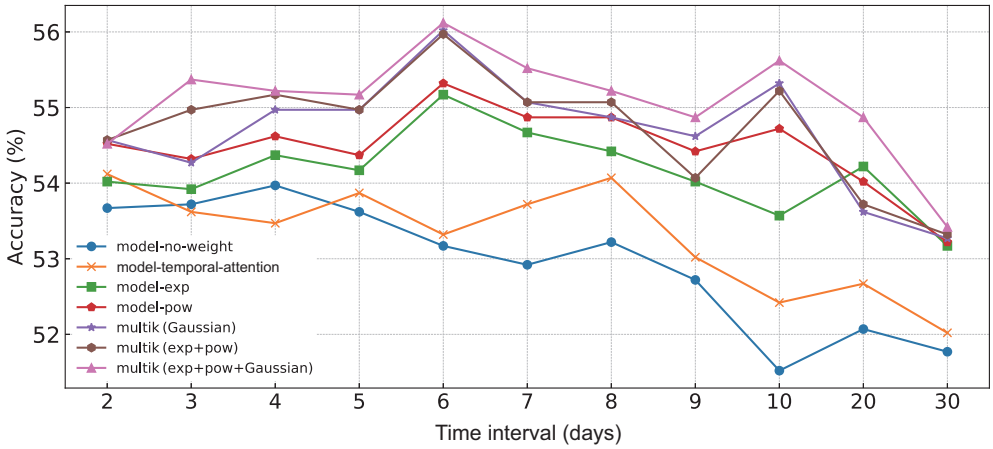

Figure 4 (Color online) Accuracy of prediction with different time intervals (from 2 days to 30 days). The Hawkes process-based models first rise and then fall as the time interval increases; they reach a peak when the time interval is 6 days.

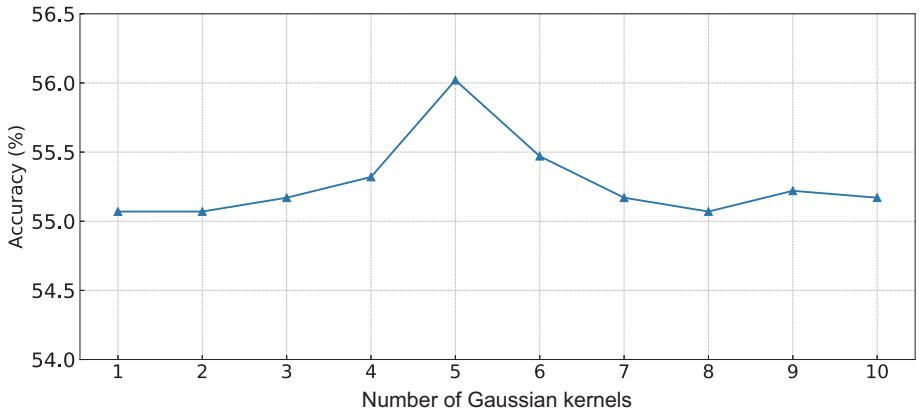

Figure 5 (Color online) Accuracy of prediction with different number of Gaussian kernels (from 1 to 10). Hawkes process based model first rises and then falls as the number of Gaussian kernels increases. They reach a peak when the number of Gaussian kernels is 5.

(1) The curves that denote the Hawkes process-based methods first rise and then fall as the time interval increases. When the time interval increases to 6 days, they reach a peak for two major reasons. First, more days of news can bring more noise information, thus limiting the performance of our approach. Second, the stock market is sensitive to information and responds quickly. More than 6 days of bad news or good news have already been priced in and no longer affect the stock market. Six days are thus a long-term time span for the stock market.

(2) The model with no-weight performs poorly with the increase in time intervals because giving equal weights to all historical financial news without considering their time-decaying influence is unreasonable. This result indicates the importance of considering the timeliness of financial news.

(3) The accuracies of our proposed multikernel Hawkes process (exp + pow + Gaussian)-based approach are consistent in different time intervals compared with the single-kernel Hawkes process. This result verifies that the multikernel function can enhance the robustness of the Hawkes process.

4.6 Accuracy vs. different number of Gaussian kernels

The number of Gaussian kernels is a hyperparameters in our model. We show the effect of different numbers of Gaussian kernels on the test set in Figure 5. The accuracy of prediction with different numbers of Gaussian kernels (from 1 to 10) is observed. The Hawkes process based model first rises and then falls as the number of Gaussian kernels increases. They reach a peak when the number of Gaussian kernels is five. When we use five Gaussian kernels, the weights of different kernels are 0.29 for the exponential kernel, 0.11 for the power-law kernel, and 0.12 for each Gaussian kernel. The exponential kernel is given the greatest weight, indicating that the time-decaying influence of long-term news is

Table 3 Return compared with different methods[a)]

Method	Return (%)
AMEX composite index	−21.65
NYSE composite index	−9.59
NASDAQ composite index	2.16
short-term	−34.81
model-no-weight	−28.81
model-temporal-attention (Hu et al. [2])	−18.82
multik (exp+pow+Gaussian) (ours)	**8.82**

a) The best result is in bold. We choose AMEX, NYSE and NASDAQ composite index as baselines because individual firms in our datasets come from these indexes.

similar to the exponential kernel, and other kernels can contribute to model the time-decaying influence of long-term news.

4.7 Market trading simulation

A shortcoming of news-driven stock prediction is that it highly depends on whether financial news of a specific firm is available on the day before the trading day. If no news is available on the previous day, short-term news-based methods cannot yield results. However, our approach can still provide a reasonable prediction result by using long-term historical news, indicating that we can conduct market transactions on more trading days compared with prediction methods based on short-term news. Thus, we propose another evaluation method, i.e., calculating the profit return in a market trading simulation.

We conduct a market trading simulation during the span of the test set by following the strategy proposed by Lavrenko et al. [27]; this strategy mimics the behavior of a daily trader who uses our model in a simple manner. If the model indicates that the most likely stock movements of specific firms will rise the next trading day, then the fictitious trader will invest; otherwise, the fictitious trader will sell short. Our model gives each stock a score based on the probability to have a rising trend if the prediction trend is rising; otherwise, the score based on the probability to have a declining trend will be given. Based on these scores, we select the top three stocks with the highest scores to construct a new portfolio for the next trading day. The selected three stocks are evenly invested at the opening price of the next trading day. After the purchase, the virtual trader holds them for one day. At the end of the day, the virtual trader sells them at the closing price. To approximate actual trading, we consider a transaction cost of 0.3% for each trading. We also respectively calculate the return by investing in the AMEX, NYSE and NASDAQ composite indices during the same period. These three returns of the composite index can not only be considered baselines but can also indicate the stock market trend. We choose the top three stocks of prediction results for the following two reasons: First, the highest profits should be obtained by selecting the topmost stocks. Second, investors always choose multiple stocks to avoid risks in case of a plunge in stock prices.

Table 3 shows the return of different methods in a market trading simulation, demonstrating that our proposed multikernel Hawkes process (exp + pow + Gaussian)-based approach can gain more profits in comparison with other methods and even performs better than the NASDAQ composite index, which outperforms all composite indices. The return rate of our proposed model is 8.82%, indicating that we can make a considerable profit in comparison with other methods. We can also observe that long-term news-based prediction methods perform better than short-term ones because they conduct transactions on more trading days, whereas short-term ones cannot. Compared with our approach, the other two long-term news-based prediction methods obtain a negative return because model-no-weight unreasonably gives an equal weight to each long-term news and model-temporal-attention cannot give a precise time-decaying influence of long-term news with the change of time. This result is due to high transaction cost, which offsets the profits. However, our method can still make profits in the downturn of the stock market, indicating that it can predict stocks with high return accurately.

5 Related work

5.1 Text-driven stock market prediction

Previous researches [28–31] leveraged time-series data to predict future movements on stock market by

utilizing historical price and column movements. The autoregressive method is mostly used in time-series stock prediction. Li et al. [32] investigated the relation between index return and contemporaneous trading volume in seven major advanced economies. With the prevalence of deep learning, a hybrid model combining recurrent neural network with autoregressive model was proposed to predict stock returns [33]. However, these methods ignore one key factor that influences the volatility of the stock market, that is, financial news.

Substantial research has found that financial news has a considerable effect on stock markets [34, 35]. Such studies investigated various text representation methods to predict the movements of stock markets [36]. Earlier approaches focused on feature engineering [27, 37], mainly utilizing n-grams as features. To augment text features, Schumaker et al. [38] extracted noun phrases and named entities through the aid of lexical and syntactic rules, as well as semantic tagging methods. Furthermore, Xie et al. [4] and Wang et al. [3] applied a well-designed semantic frame to predict stock price movement, which can be regarded as a remedy to sparse feature space problems. Sun et al. [34] explored a stock trading network beyond stock price time series and financial news for stock movement prediction. However, these methods ignore modeling events that contain considerable structure information. In an early attempt to model events, Feldman et al. [39] predefined nine categories of event types and extracted events on the basis of weighted context-free grammar. However, the scalability of this work is not strong owing to limited categories. With the advancement of open information extraction [40, 41] and deep learning methods, Ding et al. [5] represented financial news with structured events (e.g., action, actor that conducts the action, object on which the action is performed). Furthermore Ding et al. [6] used neural tensor networks to learn dense event vectors for stock market prediction.

This line of work mainly uses the title of financial news because titles are informative and contain minimal noise. To incorporate additional information on financial news documents, Duan et al. [9], Hu et al. [2], and Xu et al. [7] used an attention mechanism to learn document representations for stock movement prediction. However, previous studies mainly focused on modeling short-term effects of financial news and suffered from the weak ability of quantifying the time-decaying influence of financial news. In this study, we propose using the Hawkes process to quantify the time-decaying influence of long-term historical financial news. Modeling the influence of long-term news can consider additional information, which is beneficial for making decisions on stock market prediction. Moreover, through the quantification of the time-decaying influence, our model can distinguish financial news effects on different days with the aid of the time factor.

5.2 Hawkes process

The temporal point process is often used for modeling information cascades, a series of events that occurs in continuous time. The simplest form of temporal point process assumes that events occur independently, which is insufficient when a subsequent event is dependent on past events. Through the accumulation mechanism and time-decaying quantification, the Hawkes process [8] considers past events to predict which type of subsequent event occurs. Gao et al. [42] proposed a mixture process to model and predict retweeting dynamics, with each subprocess capturing the retweeting dynamics initiated by a key node. Cao et al. [14] proposed DeepHawkes to leverage end-to-end deep learning to make an analogy of the interpretable factors of the Hawkes process, a widely used generative process to model information cascades. In text-driven stock market prediction, financial news becomes volatile toward stock markets; as time goes by, the influence of financial news decays. Thus, we utilize this method to predict the cumulative abnormal return of individual firm by considering financial news during a period.

Lima et al. [10] showed that choosing an appropriate kernel function is crucial because different kernel choices lead to different time-decaying rates. Ogata et al. [11] utilized the power-law kernel function to model the time-decaying influence at a slow rate in earthquake and aftershock prediction. Etesami et al. [23] adopted the exponential kernel to model the time-decaying influence at a fast rate. In this study, we propose quantifying the time-decaying influence of financial news by using a multikernel function because experimental results are sensitive to the choice of kernel functions in the Hawkes process. The multikernel Hawkes process can maximize the strengths of different kernel functions to model complex variable distribution. Moreover, experimental results show that the multikernel Hawkes process can enhance the robustness of our prediction model.

6 Conclusion

In this study, we investigated the quantification of the time-decaying influence of long-term financial news using the Hawkes process for stock movement prediction. An optimal multikernel strategy for the Hawkes process further improved the robustness of our proposed model. Extensive experiments conducted on 515 listed companies demonstrated that our proposed method could yield state-of-the-art performance on stock market prediction against baseline methods. Experimental results on the market trading simulation showed that we could gain a considerable profit, as compared with the baseline methods and several composite indices. The results demonstrate that quantifying the time-decaying influence of long-term financial news is beneficial for market trading.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2018AAA0101901) and National Natural Science Foundation of China (Grant Nos. 61976073, 61702137). We thank the anonymous reviewers for their constructive comments.

References

1 Ding X, Zhang Y, Liu T, et al. Knowledge-driven event embedding for stock prediction. In: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016
2 Hu Z, Liu W, Bian J, et al. Listening to chaotic whispers: a deep learning framework for news-oriented stock trend prediction. In: Proceedings of International Conference on Web Search and Data Mining, 2018. 261–269
3 Wang W Y, Hua Z. A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014. 1155–1165
4 Xie B, Passonneau R, Wu L, et al. Semantic frames to predict stock price movement. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013. 873–883
5 Ding X, Zhang Y, Liu T, et al. Using structured events to predict stock price movement: an empirical investigation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014. 1415–1425
6 Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015. 2327–2333
7 Xu Y, Cohen S B. Stock movement prediction from tweets and historical prices. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, 2018. 1970–1979
8 Hawkes A G. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 1971, 58: 83–90
9 Duan J, Zhang Y, Ding X, et al. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In: Proceedings of International Conference on Computational Linguistics, 2018
10 Lima R, Choi J. Hawkes process kernel structure parametric search with renormalization factors. 2018. ArXiv:1805.09570
11 Ogata Y. Seismicity analysis through point-process modeling: a review. In: Proceedings of Seismicity Patterns, Their Statistical Significance and Physical Meaning, 1999. 471–507
12 Mishra S, Rizoiu M A, Xie L. Feature driven and point process approaches for popularity prediction. In: Proceedings of the Conference on Information and Knowledge Management, 2016. 1069–1078
13 Gupta A, Farajtabar M, Dilkina B, et al. Discrete interventions in hawkes processes with applications in invasive species management. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 3385–3392
14 Cao Q, Shen H, Cen K, et al. DeepHawkes: bridging the gap between prediction and understanding of information cascades. In: Proceedings of the Conference on Information and Knowledge Management, 2017. 1149–1158
15 Du N, Dai H, Trivedi R, et al. Recurrent marked temporal point processes: embedding event history to vector. In: Proceedings of KDD, 2016
16 Duan J, Ding X, Liu T. Learning sentence representations over tree structures for target-dependent classification. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018
17 Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about entailment with neural attention. 2016. ArXiv:1509.06664
18 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
19 Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 2227–2237
20 Dozat T, Manning C D. Deep biaffine attention for neural dependency parsing. 2016. ArXiv:1611.01734
21 Li J, Luong M T, Jurafsky D. A hierarchical neural autoencoder for paragraphs and documents. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015. 1106–1115
22 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. ArXiv:1409.0473
23 Etesami J, Kiyavash N, Zhang K, et al. Learning network of multivariate hawkes processes: a time series approach. In: Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence, 2016. 162–171
24 Wilson A, Adams R. Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of International Conference on Machine Learning, 2013. 1067–1075
25 Hwang Y, Tong A, Choi J. Automatic construction of nonparametric relational regression models for multiple time series. In: Proceedings of International Conference on Machine Learning, 2016. 3030–3039
26 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations, 2015
27 Lavrenko V, Schmill M, Lawrie D, et al. Mining of concurrent text and time series. In: Proceedings of KDD-2000 Workshop on Text Mining, 2000. 37–44
28 Taylor S J, Xu X. The incremental volatility information in one million foreign exchange quotations. J Empirical Finance, 1997, 4: 317–340
29 Andersen T G, Bollerslev T. Intraday periodicity and volatility persistence in financial markets. J Empirical Finance, 1997, 4: 115–158

30 Atsalakis G S, Valavanis K P. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. Expert Syst Appl, 2009, 36: 10696–10707
31 Taylor S J. Modelling Financial Time Series. Singapore: World Scientific, 2008
32 Li L, Leng S, Yang J, et al. Stock market autoregressive dynamics: a multinational comparative study with quantile regression. Math Problems Eng, 2016, 2016: 1–15
33 Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst Appl, 2015, 42: 3234–3241
34 Sun X Q, Shen H W, Cheng X Q. Trading network predicts stock price. Sci Rep, 2015, 4: 3711
35 Chen D, Zou Y, Harimoto K, et al. Incorporating fine-grained events in stock movement prediction. In: Proceedings of the 2nd Workshop on Economics and Natural Language Processing, Hong Kong, 2019. 31–40
36 Qin Y, Yang Y. What you say and how you say it matters: predicting stock volatility using verbal and vocal cues. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, 2019. 390–401
37 Luss R, D'Aspremont A. Predicting abnormal returns from news using text classification. Quantitative Finance, 2015, 15: 999–1012
38 Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news. ACM Trans Inf Syst, 2009, 27: 1–19
39 Feldman R, Rosenfeld B, Bar-Haim R, et al. The stock sonar-sentiment analysis of stocks based on a hybrid approach. In: Proceedings of the 23rd International Association of Arson Investigators Conference, 2011
40 Etzioni O, Fader A, Christensen J, et al. Open information extraction: the second generation. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011
41 Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011. 1535–1545
42 Gao J, Shen H, Liu S, et al. Modeling and predicting retweeting dynamics via a mixture process. In: Proceedings of the 25th International Conference Companion on World Wide Web, 2016. 33–34