future internet

MDPI

Article

Trend Prediction of Event Popularity from Microblogs

Xujian Zhao * and Wei Li

School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China; vl729916@aliyun.com
* Correspondence: jasonzhaoxj@gmail.com

Abstract: Owing to rapid development of the Internet and the rise of the big data era, microblog has become the main means for people to spread and obtain information. If people can accurately predict the development trend of a microblog event, it will be of great significance for the government to carry out public relations activities on network event supervision and guide the development of microblog event reasonably for network crisis. This paper presents effective solutions to deal with trend prediction of microblog events' popularity. Firstly, by selecting the influence factors and quantifying the weight of each factor with an information entropy algorithm, the microblog event popularity is modeled. Secondly, the singular spectrum analysis is carried out to decompose and reconstruct the time series of the popularity of microblog event. Then, the box chart method is used to divide the popularity of microblog event into various trend spaces. In addition, this paper exploits the Bi-LSTM model to deal with trend prediction with a sequence to label model. Finally, the comparative experimental analysis is carried out on two real data sets crawled from Sina Weibo platform. Compared to three comparative methods, the experimental results show that our proposal improves F1-score by up to 39%.

Keywords: popularity of microblog event; information entropy model; singular spectrum analysis; Bi-LSTM

Citation: Zhao, X.; Li, W. Trend Prediction of Event Popularity from Microblogs. *Future Internet* **2021**, *13*, 220. https://doi.org/10.3390/fi13090220

Academic Editor: Luis Javier García Villalba

Received: 30 July 2021
Accepted: 23 August 2021
Published: 24 August 2021

1. Introduction

With the rapid development of the Internet and the rise of the era of big data, microblog has become the main means for people to spread and obtain information. Timely and accurate prediction of the evolution trend of microblog events can help the government accurately evaluate the development trend of microblog events and provide effective decision support for the formulation of public event guidance strategies [1]. Generally, hot events on microblog platforms are often defined as the focus of public discussion and concern, which are the concentrated embodiment of netizens' interests and emotions. When an event is exposed in social network, the upsurge of the Internet media and netizens' discussion about the event on social media will affect the popularity of the events in real time. In addition, when social users exchange information with each other, they influence and are influenced by others [2]. Thus, social networks provide a large amount of real-time and continuous data for exploring the evolution of microblog events [3]. However, due to the non-linear and multivariate characteristics of microblog data, this paper has to solve two challenging problems for microblog event popularity prediction.

(1) How can the weight of each factor's impact on the popularity of microblog events be evaluated? In the previous work on microblog events' popularity prediction, most of them take one indicator as the popularity of microblog event for prediction. However, for multivariate microblog data, univariate analysis cannot well reflect the systematic change of the trend of microblog popularity. Obviously, various factors have different impacts on the popularity of microblog events, so they should play different roles in popularity prediction modeling. Therefore, it is very necessary to weigh the influence of each factor on the event popularity.

(2) How can the future trend of nonlinear time series be predicted? The popularity evolution of microblog events tends to be a nonlinear and irregular time series. Therefore, to solve the popularity trend prediction, it is necessary to extract the trend components of popularity time series. However, statistical learning methods or traditional neural network methods have a poor prediction effect on nonlinear data. Consequently, this paper needs to design an effective prediction method to denoise the time series data and extract the different components for trend prediction.

Aiming at these issues, this paper presents an effective approach to predicting the trend of microblog events' popularity. Firstly, the popularity time series of microblog events is modeled by comprehensive weighting. Secondly, the information entropy algorithm is used to measure the effects of various factors on the popularity of microblog events. Meanwhile, aiming to explore and predict the evolution trend of microblog events' popularity, this paper transforms the changes between every two time nodes in the time series into state features. Then, the box-plot method is used to divide the popularity of the microblog event into various trend spaces. Finally, this paper utilizes the Bi-LSTM model to solve the trend prediction of microblog events' popularity by learning the long-term dependence between the time steps of popularity time series.

In summary, the following contributions are made in this paper.

(1) Aiming to deal with time series modeling for nonlinear data, this paper leverages an effective model based on series data analysis to extract trend components of popularity time series. Specifically, by selecting the influence factors and quantifying the weight of each factor with information entropy algorithm, the microblog event popularity is modeled. And then, the singular spectrum analysis is carried out to decompose and reconstruct the time series of the popularity of microblog event (Section 3).

(2) This paper exploits the learning method to deal with trend prediction with a sequence to label model. Firstly, this paper models the Bi-LSTM network using past and future data from time series. Secondly, by learning the long-term dependence between the time steps of the popularity time series, the future trend prediction of microblog events is solved. Compared with the traditional LSTM model, our proposal based on Bi-LSTM network has better prediction performance (Section 3).

(3) This paper conducts experiments on a real microblog dataset from the Sina Weibo platform. Compared with three general-purpose algorithms for popularity prediction, the experimental results show that our approach achieves the best performance compared to its competitors, which provides a new solution to the trend prediction for event evolution on microblog platforms (Section 4).

2. Related Work

Scientific research on the evolution trend of microblog events can effectively monitor the development of event popularity at all stages [4–6], which is of great significance to the supervision of network opinion. At present, existing work can be divided into the following two aspects: event propagation research and event trend prediction research.

In order to reveal the propagation mechanism and the evolution law of microblog events, the pioneering work [7] carried out feature extraction on Weibo data and developed an outlier knowledge management framework for dealing with public events. Compared with traditional methods, the graph theory-based approach performed well in modeling the interaction between users [8]. Meanwhile, it is worth mentioning that text and sentiment analysis are often used to analyze netizens' attitudes when they disseminate information. For example, the LSTM model [9] was exploited to capture the features of social contexts and can integrate them into text features. Additionally, Xu et al. [10] proposed to use convolutional neural network (CNN) combined with word2vec technology to establish emotion classification model. Moreover, among all text analysis methods, LDA model was widely considered to discovery of microblog text topics [11,12].

Regarding the studies on events evolution trend prediction, the previous work is mainly divided into dynamic model and machine learning model. In Yin's work [13], a

modified epidemic model was proposed to predict the dynamics of topic reading, which represents one indicator of event popularity. In their further work [14], considering both public contact and participation on microblog, they proposed the Susceptible-Reading-Forwarding-Immune (SRFI) model to predict the overall microblog event trend in all stages. Meanwhile, Pan et al. [15] developed a Stochastic Differential Equation (SDE) to describe the observed collective patterns of the online microblogging behavior and predict the Sina Weibo volume data. On the other side, for machine learning-based method, the BP neural network [16,17] was applied to predict the trend of microblog events in early. Aimed at the sudden and non-linear characteristics of microblog events, timely grasp of the information increment of microblog events plays a key role in measuring the event evolution trend, which can be better solved by using LSTM network model [18]. Feng et al. [19] introduced LSTM model to analyze the sequence information and complete the prediction for the number of blogs for a certain event in a period. Moreover, in Mughees's work [20], the bidirectional LSTM network model is proven to be not only capable of learning long-term dependencies between the time steps of sequence data, but also can effectively use past and future information for prediction. Consequently, thanks to the ability to process time series data, the Bi-LSTM model has been successfully applied in many time series prediction tasks [21–27]. Inspired by this idea, this paper aims to exploit the learning method based on Bi-LSTM model to deal with event trend prediction with a sequence to label model.

Generally, the popularity of microblog events is affected by many factors. However, to our knowledge, so far there are very few works towards microblog events prediction considering multidimensional influential factors to model popularity index. Meanwhile, the dynamic change of microblog event popularity is easy to be nonlinear and irregular in a period of time; nevertheless, most of the previous works validate their methods and experiment only on one dataset, which may limit the applicability of the model especially for nonlinear time series data. This paper presents effective solutions to deal with trend prediction of microblog events' popularity. More specifically, by selecting the influence factors and quantifying the weight of each factor with information entropy algorithm, the microblog event popularity is modeled. And then, the singular spectrum analysis is carried out to decompose and reconstruct the time series of the popularity of the microblog event. Finally, this paper utilizes the learning method to deal with trend prediction with a sequence to label model.

3. Methodology

In order to solve the issue of trend prediction of event popularity from microblogs, this paper presents an effective predictive algorithm based on Bi-LSTM network model. Figure 1 shows the framework of our predictive system. Specifically, the system framework has three main modules, namely, data acquisition, data processing and modeling. Here, the modeling module contains the main core idea of this paper. For the data acquisition module, firstly, this paper uses the method of simulated login to enter the microblog platform, namely, Sina Weibo platform. And then the parameters and time window for retrieval are set using the advanced search function. In the end, our proposal uses Python to write a web crawler according to the functional requirements to crawl the microblog information derived from the keywords. In the data processing stage, this paper screened the results collected by the crawler to eliminate repeated and missing items. On the basis of the first two steps, this paper designs a model framework to solve the trend prediction of microblog event popularity. Compared to existing methods, our proposal proposes to model the microblog event popularity by selecting the influence factors and quantifying the weight of each factor with information entropy algorithm in order to deal with time series modeling for nonlinear data. Meanwhile, aimed at the issue of the trend prediction of nonlinear time series, the singular spectrum analysis is carried out to decompose and reconstruct the time series of the popularity of a microblog event while a learning method to deal with trend prediction with a sequence to label model is proposed in our proposal. The details will be discussed in the following four subsections.

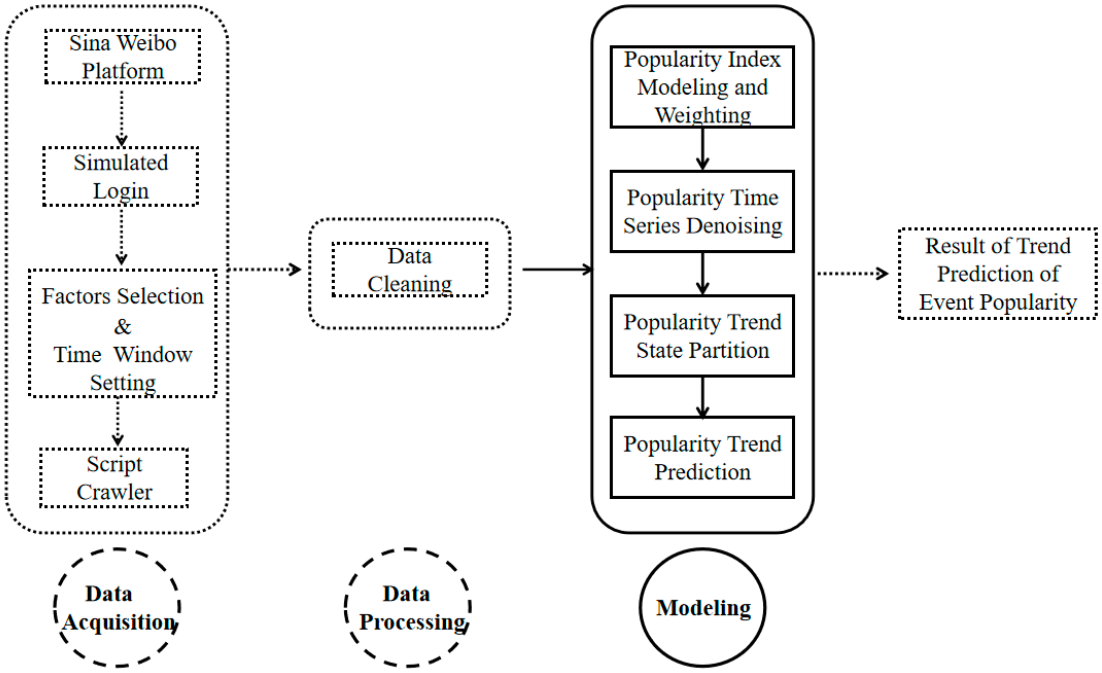


Figure 1. System framework of the event popularity trend prediction from Weibo platform.

3.1. Popularity Index Modeling and Weighting

Regarding multivariate microblog data, univariate analysis cannot well reflect the systematic change of the trend of microblog popularity. Additionally, various factors have different impacts on the popularity of microblog events, so they should play different roles in popularity prediction modeling. Meanwhile, for microblog messages, those with more user interactions have greater social popularity. Therefore, the post number, forwarding number, commenting number and the total number of likes are used to model the popularity index. In the paper, the information entropy algorithm is used to assign the weight of popularity indicators. Information entropy can be used to measure the dispersion of an index. The greater the dispersion of an index, the greater the impact of the index on the comprehensive evaluation, that is, the greater the weight.

Specifically, this paper takes the number of days of each event as the evaluation individual and the above four indicators as evaluation index. Here, in order to eliminate the dimensional differences between the evaluation indexes, this paper needs to preprocess origin data through benefit (positive) index calculation and cost (negative) index calculation, defined by Formulas (1) and (2) respectively.

$$y_{i,j} = \frac{x_{i,j} - \min(x_{i,j})}{\max(x_{i,j}) - \min(x_{i,j})} * (y_{\max} - y_{\min}) + y_{\min} \quad (1)$$

$$y_{i,j} = \frac{\max(x_{i,j}) - x_{i,j}}{\max(x_{i,j}) - \min(x_{i,j})} * (y_{\max} - y_{\min}) + y_{\min} \quad (2)$$

The symbol $\max(x_{i,j})$ and $\min(x_{i,j})$ represent the maximum and minimum of evaluation index, respectively. And then, the proportion of the evaluation object can be calculated by Formula (3).

$$p_{i,j} = \frac{y_{i,j}}{\sum y_{i,j}} \quad (3)$$

Additionally, the entropy of evaluation index is defined by Formula (4),

$$e_j = -\frac{1}{\log n} \sum p_{i,j} \log p_{i,j} \quad (4)$$

where n represents the number of evaluation index. Finally, the weight of each indicator is represented by Formula (5). Consequently, our proposal can get the time series of event popularity by index modeling and weighting.

$$w_j = \frac{1 - e_j}{\sum (1 - e_j)} \tag{5}$$

3.2. Popularity Time Series Denoising

The popularity evolution of microblog events tends to be a nonlinear and irregular time series. Therefore, to solve the popularity trend prediction, it is necessary to extract the trend components of popularity time series. Specifically, this paper needs to denoise the time series data and extract the different components for trend prediction. Our proposal decomposes and reconstructs the trajectory matrix of the time series to solve time series denoising through singular spectrum analysis [28]. Here, there are four steps for time series denoising.

(1) Embedding

Supposed that the time series of event popularity from microblog is represented by $X = [x_1, x_2, \cdots, x_m]^T$, then the paper transforms it into a p-dimensional trajectory matrix $Y = [y_1, y_2, \cdots, y_p]^T$, where $2 \leq p \leq m$. Consequently, the trajectory matrix can be defined by Formula (6).

$$Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_{m-p+1} \\ x_2 & x_3 & \cdots & x_{m-p+2} \\ \vdots & \vdots & & \vdots \\ x_p & x_{p+1} & \cdots & x_m \end{bmatrix} \tag{6}$$

(2) Decomposition

Firstly, the trajectory matrix Y obtained above is decomposed into d components, where d is the rank of matrix Y. And then, the paper can get parameter group (λ_i, U_i, V_i) of matrix YY^T by using Singular Value Decomposition (SVD) algorithm. Here, λ_i represents the singular value of the matrix, and the symbol U_i and V_i are used to define the left eigenvector and the right eigenvector, respectively. Subsequently the trajectory matrix Y and its components Y_i are defined by the following Formulas (7) and (8).

$$Y = Y_1 + Y_2 + \cdots + Y_d \tag{7}$$

$$Y_i = \sqrt{\lambda_i} U_i V_i \tag{8}$$

(3) Grouping

In the paper, k of the d components are selected as the popularity trend components, denoted as $I = \{I_1, \cdots, I_k\}$. Meanwhile, the valuable extraction component Y_I of the time series is represented by $Y_I = Y_{I_1} + Y_{I_2} + \cdots + Y_{I_k}$. Therefore, the other d-k decomposition components are considered as the noise of the time series.

(4) Reconstitution

By means of diagonal averaging, the valuable trend component Y_I formed in the grouping stage can be converted into the previous time series $x_{component} = \{x_{I_1}, \cdots, x_{I_k}\}$. Consequently, the original time series X is represented as the sum of component series data $x_{component}$ and noise series data x_{noise}.

$$X = x_{component} + x_{noise} = x_{I_1} + x_{I_2} + \cdots + x_{I_k} + x_{noise} \tag{9}$$

3.3. Popularity Trend State Partition

In order to explore and predict the popularity trend of events from microblog, this paper considers the change between each time node and the previous node into state

characteristics. Firstly, this paper needs to difference the denoised time series and calculate the state trend value $H(t)$ by Formula (10).

$$H(t) = x_{t+1} - x_t (1 \leq t \leq m-1) \tag{10}$$

However, the same data set may cover multiple hot events, and different hot events may cause different responses from netizens, which may eventually lead to a large deviation in the popularity of multiple events derived from a period of dataset. In order to explore the systematic change of event popularity, the box plot method is used to divide the popularity trend into various states. The box plot algorithm defines a standard for identifying outliers, which are usually defined as values less than $Q_L - 1.5IQR$ or greater than $Q_U + 1.5IQR$, where the symbol Q_L and Q_U represent the lower quartile and the upper quartile respectively, and the difference between Q_L and Q_U is defined as IQR. The box plot method does not require data to follow a certain distribution, so it is reasonable to use it to judge the state change of event popularity. The structure of the box plot is shown in the Figure 2.

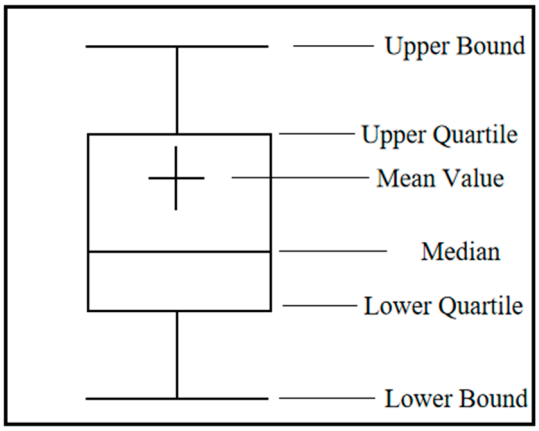


Figure 2. The structure of the box plot.

Actually, the popularity trend is divided into four states, defined by Formula (11), according to the box plot distribution.

$$\begin{aligned} S_1 &= \text{Rapid rise} = [Q_U + 1.5IQR, \overline{H}_{\max}]; \\ S_2 &= \text{Slowly rise} = [0, Q_U + 1.5IQR]; \\ S_3 &= \text{Slowly fall} = [Q_L - 1.5IQR, 0]; \\ S_4 &= \text{Rapid fall} = [\overline{H}_{\min}, Q_L - 1.5IQR]; \end{aligned} \tag{11}$$

In the paper, the lower bound and the upper bound of the box plot are set to $Q_L - 1.5IQR$ and $Q_U + 1.5IQR$, respectively. And the paper uses $\overline{H}_{\max}$ and $\overline{H}_{\min}$ to represent the maximum and minimum values of event popularity trend correspondingly.

3.4. Popularity Trend Prediction

Firstly, before modeling Bi-LSTM network for prediction, this paper needs to build the basic model, LSTM network, in order to solve sequence-to-label modeling for microblog data. In memory block of LSTM network model, there are one or more self-connected memory blocks and three multiplication units: input unit, output unit and forgetting unit. The input unit is mainly used to store the current information while the output unit is used to output the state trend changes of microblog events. Meanwhile, the forgetting unit is designed to filter valuable information and selectively forget certain past information. The network structure of LSTM model is shown as Figure 3.

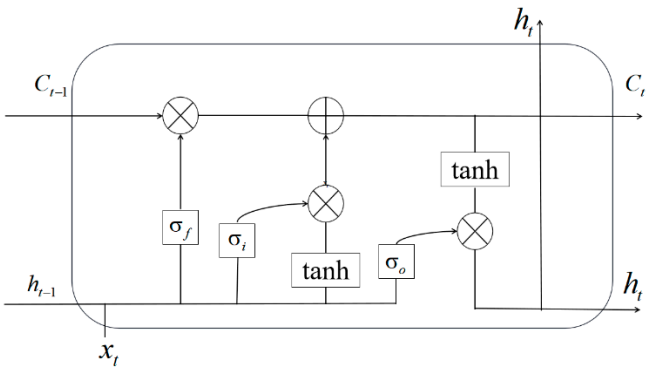


Figure 3. The network structure of LSTM model.

At each time node, the event popularity is considered as the input sequence, and the state trend data is represented as the output. Meanwhile, the input unit, output unit and forgetting unit correspond to i_t, o_t, f_t respectively. Specifically, the forgetting unit takes the current output and the previous hidden-layer-state output as input, and uses sigmoid activation function to control the unit to discard redundant information. Finally, the value of state trend, between 0 and 1, can be calculated by Formula (12).

$$f_t = \delta_g(w_f x_t + U_f h_{t-1} + b_f) \tag{12}$$

In Formula (12), the weight matrix that maps the hidden layer input to the forgetting unit is denoted as w_f, and U_f represents the weight matrix that connects the output state of the previous time node to the forgetting unit. Additionally, this paper uses the symbol b_f and δ_g to denote the offset vector and activation function, respectively.

After the information passes through the forgetting structure, the LSTM model will consider which new information to add to the unit. The added information is jointly controlled by the hidden layer state output of the previous time node h_{t-1} and the current input x_t. Through the activation function *tanh*, the new state output $\widetilde{C}_t$ is obtained. And the weight between 0 and 1 is added to each component of $\widetilde{C}_t$ to control the amount of new added information by Formulas (13) and (14).

$$i_t = \sigma_g(W_i x_i + U_i h_{t-1} + b_i) \tag{13}$$

$$\widetilde{C}_t = \tanh(W_c x_t + U_C h_{t-1} + b_c) \tag{14}$$

W_i and W_C refer to weight matrices, which map the hidden layer input to the traditional input unit and input unit states, respectively. U_i and U_C refer to weight matrices, connected to the previous unit, which map the output states to the input unit and input unit states. And the deviation vectors are defined as b_i and b_C. In addition, when information is forgotten by forgetting unit, the new unit state is updated by Formula (15).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{15}$$

Finally, the state of the output unit is calculated according to the following two Formulas (16) and (17).

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{16}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{17}$$

Here W_o is the weight matrix of the hidden layer to the output unit and U_o is the weight matrix of the output state of previous unit to the output unit. Meanwhile, this paper uses the symbol b_o to represent the deviation matrix.

Based on LSTM modeling structure, this paper aims to construct the Bi-LSTM network to explore and predict the hidden layer state of popularity time series. More specifically,

forward LSTM and reverse LSTM can obtain the past information and future information in time series, respectively. The following three Formulas (18)–(20) can reflect the mechanism of Bi-LSTM model in our work.

$$\overrightarrow{h}_t = \overrightarrow{LSTM}(h_{t-1}, x_t, C_{t-1}), t \in [1, T] \tag{18}$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(h_{t+1}, x_t, C_{t+1}), t \in [T, 1] \tag{19}$$

$$H_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{20}$$

The symbol T refers to the time span of microblog event. And the structure of Bi-LSTM network is shown in Figure 4.

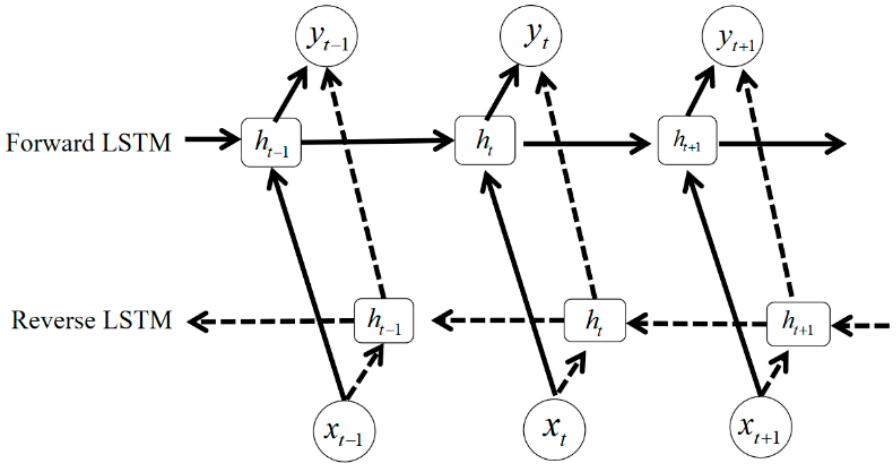


Figure 4. Bi-LSTM network structure based on LSTM.

4. Experiment

In this section, the performance evaluation of our algorithms is discussed. This paper aims to measure the effectiveness of our proposal for trend prediction of event popularity from microblogs.

4.1. Dataset

This paper uses the crawler to collect data by means of daily statistics with "genetically modified food" as the search keyword from Sina Weibo platform. After eliminating advertisements, repeated posts and irrelevant microblogs, a total of 28,326 Weibo messages from 1 January 2017 to 9 July 2019 are retrieved from microblogs. The details of the Weibo posts are as shown in Table 1.

Table 1. The statistical results of Weibo posts.

Indicator	Minimum	Maximum	Median Value	Standard Deviation
post number	0	561	40.42	402.09
forwarding number	0	44,547	781.81	1.82
commenting number	0	6609	83.23	43.21
number of likes	0	139,130	1344.20	9.87

In order to verify the effectiveness and universality of our approach, the paper selects two representative events from Weibo posts as datasets to build the predictive model. The specific datasets are as shown in Table 2.

Table 2. The details of event datasets.

	Time Span	Event
Dataset 1	14 July 2017–3 November 2017	CCTV news microblog refutes ten incidents of genetically modified food
Dataset 2	4 January 2019–23 May 2019	The famous blogger posts a blog: the harm of genetically modified food will be written into the textbook of colleges

Table 3 gives a sample of dataset 1, which has four popularity factors and spans three days.

Table 3. The details of sample data in dataset 1.

Time	Popularity Factors			
	#Post Number	#Forwarding	#Commenting	#Likes
14 July 2017	21	75	5	20
15 July 2017	7	32	0	50
16 July 2017	7	73	5	24

In order to train the model, this paper needs to divide the data set into training set and test set. In this paper, the model is trained by about 83% of the time series length, and evaluate the performance of the model by about 17%. More specifically, the paper takes the microblog messages of 94 days from 4 July 2017 to 15 October 2017 in dataset 1 as the training set, and the remaining 19 days as the test set. In data set 2, a total of 116 days of microblog messages from 4 January 2019 to 29 April 2017 are used as the training set, and the remaining 24 days of posts are used as the test set.

4.2. Evaluation Metrics

Aiming to evaluate the prediction accuracy of the proposed model, this paper performs model evaluations in terms of different metrics. More specifically, the evaluation metrics are defined as follows:

(1) Precision

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{21}$$

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \tag{22}$$

$$P_{weight} = \sum_{i=1}^{n} (P_i \cdot w_i) \tag{23}$$

(2) Recall

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{24}$$

$$R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \tag{25}$$

$$R_{weight} = \sum_{i=1}^{n} (R_i \cdot w_i) \tag{26}$$

(3) F1-score

$$F_i = 2\frac{P_i \cdot R_i}{P_i + R_i} \tag{27}$$

$$F_{micro} = 2\frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \tag{28}$$

$$F_{weight} = \sum_{i=1}^{n} (F_i \cdot w_i) \tag{29}$$

In the paper, *TP* represents the number of samples that are labeled as positive samples and also classified as positive samples; *FP* refers to the number of samples labeled as negative samples but classified as positive samples; *FN* is the number of samples that are labeled as positive samples and classified as negative samples. However, this paper mainly focuses on the prediction accuracy of the popularity trend of microblog events. Therefore, this paper chooses specific three metrics that pay more attention to the accuracy to evaluate the prediction model, which are P_{micro}, P_{weight} and F_{weight}.

4.3. Results of Event Popularity Time Series Construction

This paper conducts experiments for popularity index modeling and weighting on two datasets. Specifically, according to Formulas (1)–(5) in Section 3.1, the paper uses the information entropy model to measure the microblog event popularity, and then construct the time series of event popularity, as shown in Figure 5.

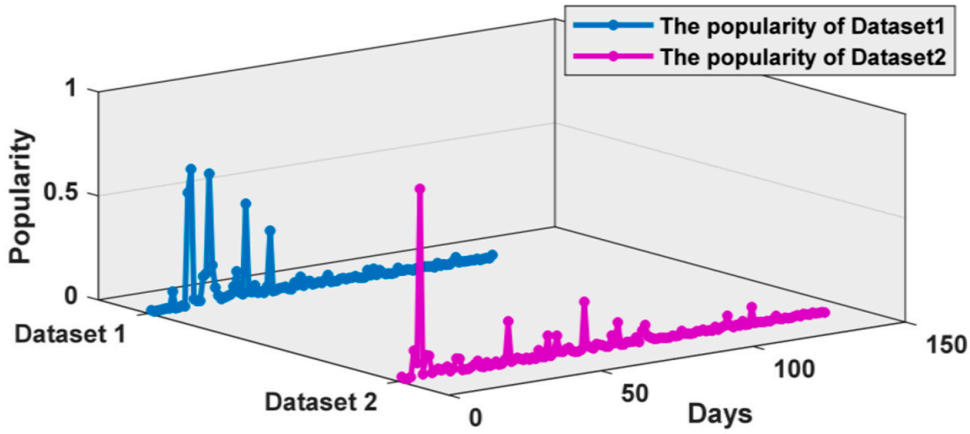


Figure 5. Time series of microblog events on the topic of genetically modified food.

4.4. Results of Popularity Time Series Denoising

According to the observed changes of time series on the two data sets in Figure 5, it is found that the two time series have the characteristics of chaos and nonlinearity. Therefore, the trend components of the two time-series need to be extracted to effectively eliminate the further influence of noise on the subsequent prediction modeling and improve the prediction accuracy.

Firstly, the data set is divided into training set and test set. And then, the singular spectrum analysis method is applied into the two data sets respectively on the training set. After component extraction and series denoising according to Formulas (6)–(9) in Section 3.2, the new time series is reconstructed. The differences between the original data and the denoised data are shown in Figure 6.

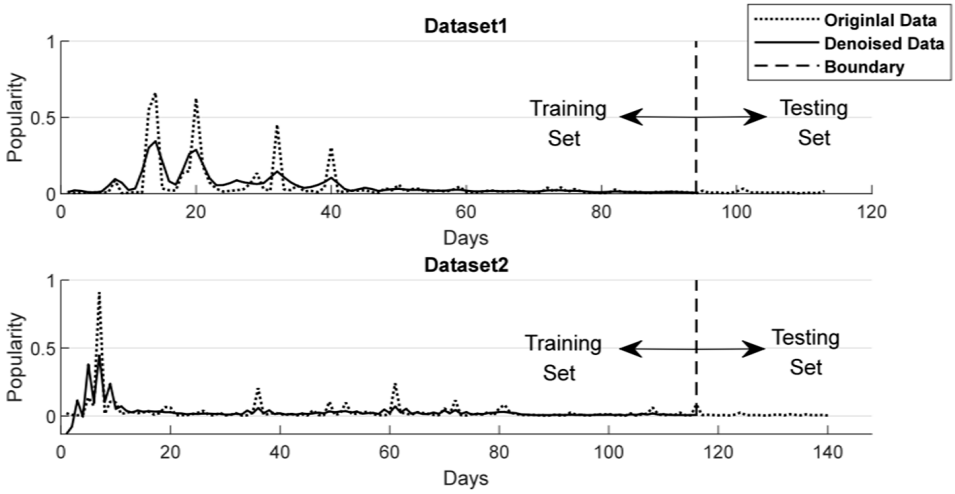
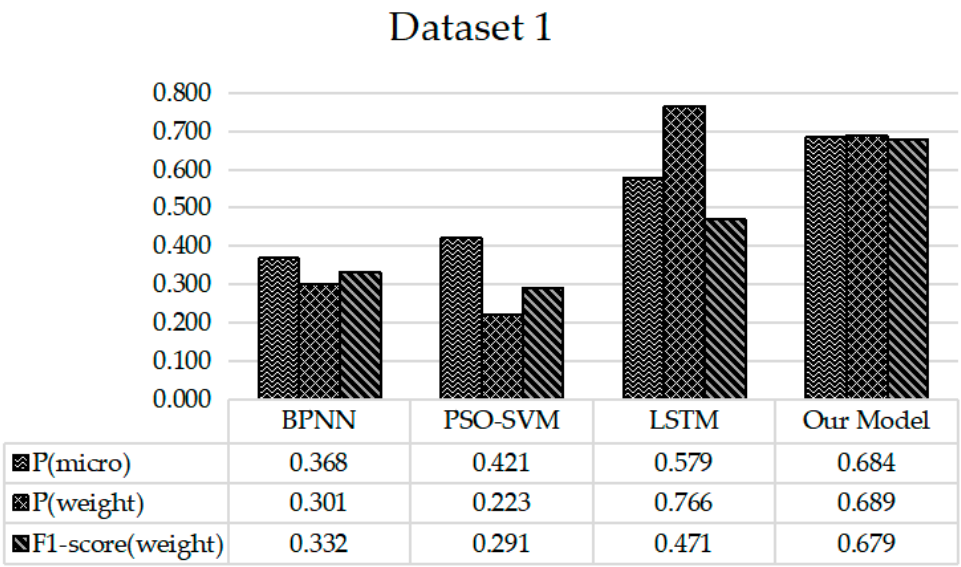


Figure 6. Comparison of popularity time series before and after denoising.

4.5. Results of Popularity Trend Prediction

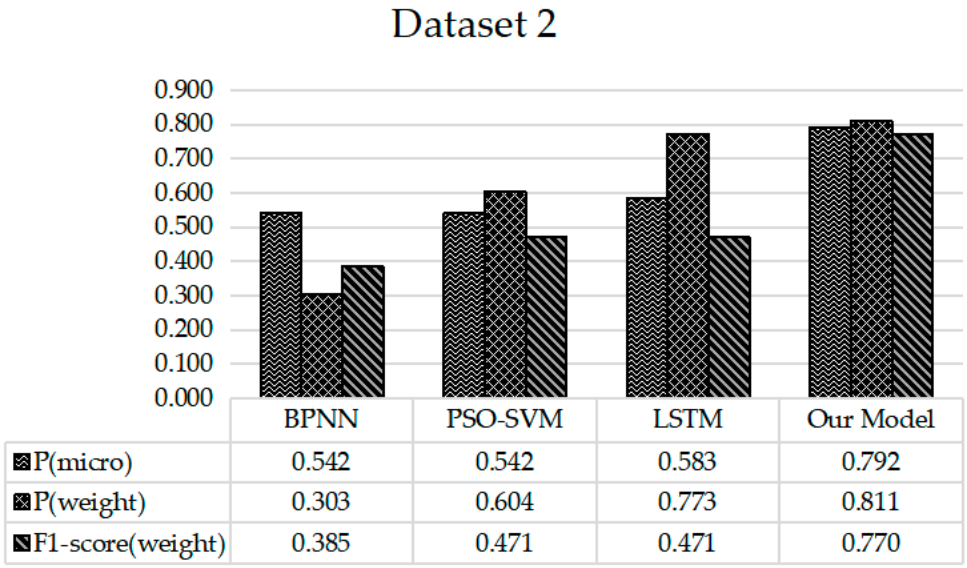
This paper takes the popularity of microblog events as input and the changed state, measured by box plot algorithm, as output to build Bi-LSTM network model. For the evaluation of predictive performance, the paper compares our proposal with three existing methods, including BP neural network (BPNN) [16], Particle Swarm Optimization Based Support Vector Machine model (PSO-SVM) [29] and Long Short Term Memory network (LSTM) [18]. Meanwhile, the model effectiveness is measured in terms of the metric P_{micro}, P_{weight} and F_{weight}.

The experimental results as shown in Figures 7 and 8 show that our model achieves better results compared with other algorithms in the task of popularity trend prediction. Especially, compared with the basic LSTM network, the proposed model in the paper is characterized by component extraction on the basis of LSTM to achieve the effect of denoising. In addition, considering the influence of future information on event popularity, the model forms a bidirectional LSTM, so it has better performance than the basic LSTM model.



	BPNN	PSO-SVM	LSTM	Our Model
P(micro)	0.368	0.421	0.579	0.684
P(weight)	0.301	0.223	0.766	0.689
F1-score(weight)	0.332	0.291	0.471	0.679

Figure 7. Results of popularity trend prediction on dataset 1.



	BPNN	PSO-SVM	LSTM	Our Model
P(micro)	0.542	0.542	0.583	0.792
P(weight)	0.303	0.604	0.773	0.811
F1-score(weight)	0.385	0.471	0.471	0.770

Figure 8. Results of popularity trend prediction on dataset 2.

5. Conclusions

This paper proposes an effective approach to deal with popularity trend prediction for microblog events based on SSA and Bi-LSTM model. Firstly, the algorithm of singular spectrum analysis is used to extract trend components of popularity time series from Weibo posts. Then, after time series denoising by SSA model, the event popularity is divided into different trend states by using the box plot analysis. Finally, the paper exploits the Bi-LSTM model to deal with popularity trend prediction with a sequence to label model. Meanwhile, the comparative experiments on two real datasets with three existing methods are conducted. The experimental results show that our model performs best on both datasets with respect to various metrics, which demonstrates the superiority of our proposal. In the future, this paper will explore how to follow some KDD standards, e.g., CRISP-DM standard, to optimize the process of the system. Meanwhile, the other language datasets will also be considered to verify and improve the system performance on cross-language.

Author Contributions: Conceptualization, X.Z. and W.L.; methodology, X.Z. and W.L.; data curation, W.L.; writing—original draft preparation, X.Z. and W.L.; writing—review and editing, X.Z. and W.L.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Humanities and Social Sciences Foundation of the Ministry of Education, grant number 17YJCZH260, and the CERNET Innovation Project, grant number NGII20180403, and the Sichuan Science and Technology Program, China, grant number 2020YFS0057.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fang, S.W.; Zhao, N.; Chen, N.; Xiong, F.; Yi, Y.H. Analyzing and predicting network public opinion evolution based on group persuasion force of populism. *Phys. A Stat. Mech. Its Appl.* **2019**, *525*, 809–824. [CrossRef]
2. Mao, Y.B.; Bolouki, S.; Akyol, E. On the Evolution of Public Opinion in the Presence of Confirmation BIAS. In Proceedings of the 2018 IEEE Conference on Decision and Control (CDC), Miami, FL, USA, 17–19 December 2018.
3. Wang, F.; Li, H.T.; Zuo, J.; Wang, Z. Evolution of online public opinions on social impact induced by NIMBY facility. *Environ. Impact Assess. Rev.* **2019**, *78*, 106290. [CrossRef]
4. Mu, L.; Jin, P.Q.; Zhao, J.; Chen, E.H. Detecting evolutionary stages of events on social media: A graph-kernel-based approach. *Future Gener. Comput. Syst.* **2021**, *123*, 219–232. [CrossRef]
5. Mu, L.; Jin, P.Q.; Zheng, L.Z.; Chen, E.H. EventSys: Tracking Event Evolution on Microblogging Platforms. In Proceedings of the 23rd International Conference on Database Systems for Advanced Applications, Gold Coast, QLD, Australia, 21–24 May 2018.

6. Mu, L.; Jin, P.Q.; Zheng, L.Z.; Chen, E.H.; Yue, L.H. Lifecycle-Based Event Detection from Microblogs. In Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018.
7. Xia, H.S.; An, W.Y.; Li, J.Z.; Zhang, Z.P. Outlier knowledge management for extreme public health events: Understanding public opinions about COVID-19 based on microblog data. *Socio-Econ. Plan. Sci.* **2020**, 100941. [CrossRef]
8. Xiong, J.; Feng, X.D.; Tang, Z.W. Understanding user-to-User interaction on government microblogs: An exponential random graph model with the homophily and emotional effect. *Inf. Process. Manag.* **2020**, *57*, 102229. [CrossRef]
9. Yang, J.; Zou, X.M.; Zhang, W.; Han, H.Y. Microblog sentiment analysis via embedding social contexts into an attentive LSTM. *Eng. Appl. Artif. Intell.* **2021**, *97*, 104048. [CrossRef]
10. Xu, D.L.; Tian, Z.H.; Lai, R.F.; Kong, X.T.; Tan, Z.Y.; Shi, W. Deep learning based emotion analysis of microblog texts. *Inf. Fusion* **2020**, *64*, 1–11. [CrossRef]
11. Hajjem, M.; Latiri, C. Combining IR and LDA Topic Modeling for Filtering Microblogs. *Procedia Comput. Sci.* **2017**, *112*, 761–770. [CrossRef]
12. Xu, F.; Sheng, V.S.; Wang, M.W. Near real-time topic-driven rumor detection in source microblogs. *Knowl.-Based Syst.* **2020**, *207*, 106391. [CrossRef]
13. Yin, F.L.; Wu, J.L.; Shao, X.Y.; Wu, J.H. Topic reading dynamics of the Chinese Sina-Microblog. *Chaos Solitons Fractals X* **2020**, *5*, 100031. [CrossRef]
14. Yin, F.L.; Pang, H.Y.; Xia, X.Y.; Shao, X.Y.; Wu, J.H. COVID-19 information contact and participation analysis and dynamic prediction in the Chinese Sina-microblog. *Phys. A Stat. Mech. Its Appl.* **2021**, *570*, 125788. [CrossRef]
15. Pan, J.S.; Li, Y.Q.; Hu, H.P.; Hu, Y. Modeling collective behavior of posting microblogs by stochastic differential equation with jump. *Phys. A Stat. Mech. Its Appl.* **2021**, *578*, 126117. [CrossRef]
16. Zeng, Z.M.; Huang, C.Y. Research on public opinion heat trend prediction model of emergent infectious diseases based on BP neural network. *J. Mod. Inf.* **2018**, *38*, 37–52.
17. Huang, Y.J.; Chen, F.J.; You, D.D. Research on prediction of network public opinion based on hybrid algorithm and BP neural network. *Inf. Sci.* **2018**, *36*, 24–29.
18. Jing, N.; Hu, Y.; Han, X.S. Trend of COVID-19 network attention based on ARIMA and LSTM. *China Saf. Sci. J.* **2020**, *12*, 37–42.
19. Feng, Y.; Lv, H.X.; Xu, H.Y.; Wang, R.B.; Zhang, Y.G. The Network Trend Prediction Model of Public Opinion Events Based on SDZ-LSTM. *Inf. Stud. Theory Appl.* **2021**, *44*, 158–163.
20. Mughees, N.; Mohsin, S.A.; Mughees, A.; Mughees, A. Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. *Expert Syst. Appl.* **2021**, *175*, 114844. [CrossRef]
21. Liang, T.; Zhao, Q.; Lv, Q.Z.; Sun, H.X. A novel wind speed prediction strategy based on Bi-LSTM, MOOFADA and transfer learning for centralized control centers. *Energy* **2021**, *230*, 120904. [CrossRef]
22. Shahid, F.; Zameer, A.; Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* **2020**, *140*, 110212. [CrossRef] [PubMed]
23. Zhang, B.; Zhang, H.W.; Zhao, G.M.; Lian, J. Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environ. Model. Softw.* **2020**, *124*, 104600. [CrossRef]
24. Özkaya, U.; Öztürk, Ş.; Melgani, F.; Seyfi, L. Residual CNN + Bi-LSTM model to analyze GPR B scan images. *Autom. Constr.* **2021**, *123*, 103525. [CrossRef]
25. Wang, Y.; Zhang, M.; Wu, R.M.; Wang, H.Y.; Luo, Z.Y.; Li, G. Speech neuromuscular decoding based on spectrogram images using conformal predictors with Bi-LSTM. *Neurocomputing* **2021**, *451*, 25–34. [CrossRef]
26. Li, D.D.; Liu, J.L.; Yang, Z.; Sun, L.Y.; Wang, Z. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Syst. Appl.* **2021**, *173*, 114683. [CrossRef]
27. Zhu, Z.J.; Dai, W.H.; Hu, Y.; Li, J.S. Speech emotion recognition model based on Bi-GRU and Focal Loss. *Pattern Recognit. Lett.* **2020**, *140*, 358–365. [CrossRef]
28. Mi, X.W.; Liu, H.; Li, Y.F. Wind speed prediction model using singular spectrum analysis, empirical mode decomposition and convolutional support vector machine. *Energy Convers. Manag.* **2019**, *180*, 196–205. [CrossRef]
29. Wu, Q.; Yan, H.S.; Yang, H.B. A Forecasting Model Based Support Vector Machine and Particle Swarm Optimization. In Proceedings of the 2008 Workshop on Power Electronics and Intelligent Transportation System, Guangzhou, China, 2–3 August 2008.