Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Early Warning Scheme of COVID-19 related Internet Public Opinion based on RVM-L Model

Rongbo Zhu a,*, Qianao Ding b, Mai Yu b, Jun Wang b, Maode Ma c

a College of Informatics, Huazhong Agricultural University, Wuhan, China
b College of Computer Science, South-Central University for Nationalities, Wuhan, China
c School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

1. Introduction

Public opinion monitoring is a key issue in modern society, as it can provide opportune information for analyzing and predicting the trends and patterns of unexpected events on the Internet to prevent social conflicts [1, 2, 3]. In recent years, online information on emergencies has attracted great public attention. For example, throughout the development of the coronavirus disease 2019 (COVID-19) pandemic [4, 5, 6], the scale and speed of information diffusion have rapidly increased, and the focus of public opinion monitoring has shifted from tracking to early warning [7, 8]. Information that becomes a popular concern induces public anxiety and panic, which has greatly hurt the enthusiasm of the people to fight the pandemic [9, 10]. If relevant departments can effectively monitor and analyze public opinion and generate prompt warnings to refute rumors and release official news, this will help people better understand the true trend of COVID-19, avoid unnecessary panic, and ensure normal order and social stability [11]. Therefore, achieving effective public opinion monitoring and analysis is a key issue for improving the public’s awareness of preventive measures, helping society organize management efforts, and effectively guiding the development of public opinion. When a public opinion incident breaks out, the relevant information quickly spreads in a very short time, leading to widespread concern among the masses. The ability to predict such hot events in real time can also help avoid the occurrence of secondary related events caused by public opinion.

To improve the prediction accuracy for Internet public opinion, the existing work has focused on three aspects: prediction models based on time-series data, prediction models based on artificial intelligence, and prediction models based on hybrid optimization.

Prediction models based on time-series data are suitable for most types of public opinion analysis [12]. Such models include autoregressive integrated moving average (ARIMA) models [13], autoregressive models [14], and logistic models [15] and can be effectively used to fit the factors influencing public opinion and predict future trends. For a logistic model, the prediction curve tends to show a higher degree of fit to the natural trend of public opinion growth [16]. Although logistic models are often used in multiclassification problems, saturated variable problems can also be described using logistic models [17, 18], which can also be used to predict the trend of public opinion when the amount of data available is small. However, great challenges arise in selecting the factors to represent the trend of public opinion. Because of the better prediction performance of time-series prediction models, such models are commonly used for various forecasting tasks.
such as fault prediction in industrial fields [19] and information dissemination analysis in public opinion monitoring [20]. However, public opinion trends are influenced by many factors on the Internet, and the relationships among the various factors are often complex and tightly coupled.

In the complex Internet environment, there are many different factors that can affect public opinion. Artificial intelligence-based prediction models, such as back propagation neural networks (BPNs) [21], support vector machines (SVMs) [22], Bayesian models [23], and cloud models [24], can effectively predict how trends and values are affected by multiple factors based on a large amount of historical data. However, for abruptly emerging public opinion events with only a small amount of associated data in the very short term, such schemes have difficulty effectively predicting the relevant trends since their prediction accuracy depends on the quantity and quality of historical data [25]. Thus, for “bursty” events with little information over the short term, they face some deficiencies. For SVM-based prediction models, the prediction results are closely related to the value of the penalty factor [26], while Bayesian-based schemes are very sensitive to the feature selection process, which is not conducive to predicting trends quickly [27]. In addition, such models have difficulty analyzing the key points regarding public opinion trends based on limited data, resulting in a bottleneck hindering early warning for public opinion emergencies.

In practical application scenarios, the prediction performance is affected by multiple factors and the dynamic characteristics of COVID-19-related events. To improve the prediction accuracy, a multinomial logistic model has been proposed [28, 29], in which the convergence of Bayesian-based schemes is used as the basis for a mathematical method of predicting related user behavior determines the trends of public opinion events. Considering the influence of multiple factors, and public opinion trends are influenced by many factors on the Internet, limited data and accurate prediction. Specifically, we attempt to take advantage of the capabilities of RVMs for dealing with multifactor event forecasting and the capabilities of logistic regression in dealing with overall trend prediction. The contributions of this paper are summarized as follows.

(1) A hybrid RVM and logistic regression (RVM-L) model is proposed. To improve the forecasting accuracy for burst events with insufficient data, RVM-L model incorporates multivariate analysis and adopts Lagrange interpolation to fill in the gaps in predicted trends.

(2) A novel metric called the critical interval for early warning is introduced to enhance the early warning performance for COVID-19-related events. To overcome inaccurate early warning point prediction, the critical interval captures the lower and upper bounds of trend variation for burst events.

(3) An early warning scheme is proposed, which comprehensively considers the propagation laws of multiple attributes of Internet public opinion and the dynamic characteristics of COVID-19-related burst events.

The rest of the paper is organized as follows. Section 2 introduces the proposed model and early warning scheme. Section 3 presents experimental results to verify the proposed scheme. Finally, the paper is concluded in Section 4.

2. Proposed scheme

Problem description

To capture and predict the trends of Internet public opinion, a hybrid RVM-L model is proposed. And the system framework is shown in Figure 1. To overcome the errors incurred in the case of small samples, Lagrangian interpolation is adopted. For dynamic trend forecasting, the concept of an early warning point is extended to the concept of an early warning interval to improve the reliability of prediction. The related notations are listed in Table 1.

As shown in Figure 1, the system framework consists of three parts: a data storage center, a data analysis center, and an early warning center. Based on a survey of user behaviors in response to hot events, it has been found that users most commonly use a social media platform’s forwarding, commenting, and like functions, and the accumulation of related user behavior determines the trends of public opinion events. Therefore, the data collected by the data storage center include the number of forwards, the number of comments, and the number of likes. In the data analysis center, the RVM-L model is used to predict the trends of public opinion events based on the influence of multiple factors, and Lagrangian interpolation is used to correct the predicted trends. The early warning center feeds the early warning results from the data analysis center back to users in a timely manner, providing timely and effective early warning functions.

To capture the evolutionary trends of emergencies, an information crawler collects web information and extracts the numbers of forwards, comments, and likes. An RVM model is then used to perform multivariate fitting to obtain a synthesized heat degree, and the results of the RVM analysis are, in turn, input into a logistic model to predict an overall trend that can capture the evolution of public opinion. Considering the time-varying and small-sample nature of public opinion data, Lagrangian interpolation is adopted to predict an effective value within a certain interval as a supplement to the data, and on this basis, the predicted trend is optimized to increase the confidence of the prediction. Finally, the mathematical model of the public opinion trend is used to
posed scheme is shown in Figure 2.

3

Table 1

| Symbol       | Description                                      |
|--------------|--------------------------------------------------|
| $h_{s,t}$    | Number of likes                                  |
| $h_{c,t}$    | Number of comments                               |
| $h_{f,t}$    | Number of forwarding                             |
| $x_i$        | Data set                                         |
| $H_v$        | RVM multi-factor fitting target vector           |
| $\Phi$       | Matrix containing the kernel function            |
| $\alpha$     | Weight coefficient matrix                        |
| $\varepsilon$| Noise vector representation                      |
| $\sigma$     | Growth rate                                      |
| $\gamma$     | Saturation variable                              |
| $\omega$     | Hyperparameters                                  |
| $\sum$       | Covariance                                       |
| $\nu$        | Average value                                    |
| $H_v(t)$     | Public opinion trend heat function               |
| $\delta$     | Two-point slope                                   |
| $\omega$     | Interpolation node                               |
| $\mu$        | Error remainder                                  |
| $\Phi(x)$    | The largest derivative of the function           |
| $d_0$        | Initial data volume                              |
| $d_n$        | Interpolation data volume                        |
| $\sigma^2$   | Error range                                      |
| $H^*(t)$     | Two-point slope                                   |
| $h_i(t)$     | Lagrange interpolation polynomial coefficients   |
| $\lambda$    | Critical section evaluates given range value     |
| $[t_p,t_q]$  | Early warning critical interval                  |

To describe the multiple attributes and laws governing the propagation of public opinion, an RVM model is adopted to obtain a heat degree that can represent the trend of public opinion. A logistic model then uses the synthesized heat degree to predict the trends of Internet public opinion events. The proposed hybrid RVM-L model is shown in Figure 3.

Three public opinion factors, namely, the number of likes $h_{s,t}$, the number of comments $h_{c,t}$, and the number of forwards $h_{f,t}$, are adopted in the model. Taking these factors as the input vector $x$ and the synthesized hotness value results as the target vector $H_v$, the RVM model is used to fit a synthesized hotness value for multiple public opinion factors. The relationship between these three factors and the synthesized heat degree $H_v$ can be expressed as:

$$H_v = \Phi(x)\omega + \varepsilon$$

where $\omega$ is a set of weight coefficients expressing the degrees to which the different factors contribute to the heat degree, $\omega = (\omega_1, \omega_2, \omega_3)^T$; $\Phi(x) = [1, K(x,x_1), K(x,x_2)]$ maps the nonlinear sequence of samples to a higher-dimensional feature space, in which the samples form a linear sequence; and $K(x,x_i)$ is the Gaussian kernel function of RVM. Then, the relationship between each sample and the corresponding weight coefficients $\omega$ can be expressed as:

$$p(\omega|A) = \prod_{i=0}^{N} N(\alpha_i|0, \tau_i^{-1})$$

where $A$ is a diagonal matrix containing all hyperparameters $\alpha$. The noise vector $\varepsilon$ obeys a zero-mean Gaussian distribution with a variance of $\sigma^2$. Then, once the dependency between the input vector and the target vector has been determined, the maximum likelihood estimate of the data can be obtained as:

$$P(\{t_{n,0}, \omega, \sigma^2\}) = \prod_{n=1}^{N} N(t_{n}|\Phi(x_n)\omega_n, \sigma^2)$$

According to formulas (2) and (3), the posterior stepwise probability of the weight coefficient vector $\omega$ can be obtained as:

$$p(\omega|t_{n,0}, \sigma^2, A) = N(\omega|\mu, \sum)$$
where the covariance is \( \sum = (A + \sigma^{-2} \Phi \Phi^T)^{-1} \) and the average value is \( \sum = (A + \sigma^{-2} \Phi \Phi^T)^{-1} \).

The probability distribution \( p(t, A, \sigma^2) \) of the target vector can be obtained by solving the prior distribution and the posterior distribution, and the unknown parameters \( a \) and \( \sigma^2 \) can be solved for using the method of fast marginal likelihood maximization. In this way, when a new input vector becomes available, the output target vector can be obtained, and the synthesized hotness value point pair \((t, h)\) can be generated through multifactor fitting with the RVM model.

In the data analysis center, a logistic model is applied to the synthesized heat degree data \((t, h)\) to predict the trends of Internet public opinion events. We obtain:

\[
H^*(t) = \frac{S}{1 + ae^{-at}}
\]

where the public opinion popularity function \( H^*(t) \) has a time-varying value, \( S \) is the saturation value of public opinion, and \( w \) represents the growth rate of the public opinion trend. Logistic models are often used to solve multiclassification problems, and a logistic function is also a useful way to express a saturated variable.

**Lagrangian interpolation optimization**

An RVM-L model can generate a fitted public opinion curve that is well consistent with the real public opinion trend. However, in real applications, the small samples associated with bursty events tend to result in errors in peak values. As a result, the peak value will often be too low or too high to trigger the early warning mechanism. To overcome the prediction error for small samples, Lagrangian interpolation is adopted to improve the prediction performance of the RVM-L model.

For data \( y_i \) to be fitted, the heat degree function of the first predicted value is \( H^* \). Suppose that the real public opinion heat degree function is \( H_t \), which can be obtained in accordance with the best uniform polynomial approximation. There is a closed interval \([c, d]\) in the forecast time series such that \( H_t \in C[c, d] \). Then, there is a polynomial \( P_n(t) \) such that the error satisfies the following condition:

\[
\|H_t - P_n(t)\|_\infty = \max_{t \in [c,d]} |H_t(P_n(t))| = \min_{t \in C[c, d]} |H_t - P_n(t)|
\]

The best uniform polynomial approximation is obtained when the above condition is met, and accordingly, approximate values of the original data points that do not lie in the interval \([c, d]\) can be obtained. Therefore, this approximation can be used as a supplement to the historical data for subsequent simulation using the logistic model.

In the interval \([c, d]\), let \( t = \frac{c + d}{2} + \frac{d - c}{2}\cdot x, \ x \in [-1, 1] \). According to the definition of the Chebyshev polynomials, the interpolation time \( t_i \) can be expressed as:

\[
t_i = \frac{c + d}{2} + \frac{d - c}{2} \cdot \cos\left(\frac{2k + 1}{2n + 2} \pi\right).
\]

The determination of the degree \( n \) of the polynomial corresponding to the best uniform approximation is related to the error (remainder) between the original function and the approximate polynomial. And the remainder term \( R_{n}(t) \) is:

\[
R_{n}(t) = \frac{y^{(n+1)}(\theta)}{(n+1)!} (d - c)^{n+1} \frac{T_{n+1}(t)}{2^n + 1}
\]

where \( y^{(n+1)}(\theta) \) is the maximum value of the derivative of the function. The logistic model fits an S-shaped curve, such that the model finds the second and third partial derivatives of \( t \) to identify the point where the rate of trend increase of the curve starts to rise, the point where the rate of trend increase is maximal, and the point where the rate of trend increase starts to slow down. Therefore, the following three situations are considered:

1. If the number \( d_{\text{in}} + d_{\text{d}} \) of inserted points is fewer than the number of points to reach the maximum growth rate, the value of \( \theta \) in \( y^{(n+1)}(\theta) \) is the same as the value at the right end of the interval \([c, d]\).
Given the historical data set \( \{x_i, y_i\} \), an interpolation point \( t_i \), and the last data point \( (x_d, y_d) \), we can obtain the value \( y_i \) of the next point to be inserted:

\[ y_{\text{next}} = k(x_i - x_d) + y_d \]

where \( k \) is the slope. It can be seen that the interpolation interval is closely related to the key points corresponding to changes in the trend of public opinion. To calculate the value at the next insertion point, the following two cases are considered:

1. When \( d_i > t_0 \), \( t = x_d \), we have:
   \[ k = \frac{y_{\text{next}} - y_d}{x_d - x_i} \]  

2. When \( d_i < t_0 \), the amount of data is very small, and the slope of the curve predicted by the logistic model will be much greater than the true value. To solve this problem, the interval is moved backward by one interval, and the slope \( k \) at each point is obtained from the historical data interval. That is, \( t = (x_d - d + c) \) is substituted into the first derivative of the predicted curve, and we obtain:
   \[ k = \frac{y_{\text{next}} - y_d}{x_d - x_i} \]  

Then, we can determine the degree \( n \) of the polynomial and the interpolation node \( (t, y) \). In the interpolation interval \([c, d]\), the interpolation polynomial can be obtained in accordance with the Lagrange interpolation theorem as follows:

\[ L_n(t) = \sum_{i=0}^{n} l_i(t)y_i \]

where:

\[ l_i(t) = \frac{\prod_{j=0}^{n} (t - t_j)}{\prod_{j=0}^{n} (t_i - t_j)} \]

**Critical interval for early warning**

Once the forecast trend of public opinion has been obtained, it is necessary to analyze it to solve for the key points where the public opinion trend changes. However, when the trend is changing dynamically, using static points as warning points will lead to errors in warning behavior. Therefore, the concept of an early warning point is extended to the concept of an early warning critical interval to adapt to dynamic changes in public opinion trends and achieve more accurate warning results.

Definition 3: For a known key point \( t \), given a value \( \gamma > 0 \), let the heat degree \( H^*(t) \) at the key point satisfy \( |H^*(t)| \leq \gamma \); then, the interval \([t_p, t_q] \]

\[ f'' = \frac{\text{Saw} \cdot e^{-\gamma t} (ae^{-\gamma t} - 1)}{(1 + ae^{-\gamma t})^3} \]  

\[ f' = \frac{\text{Saw} \cdot e^{-\gamma t} (3a - 2ae^{-\gamma t} + 1)}{(1 + ae^{-\gamma t})^3} \]

By setting \( f'' = 0 \) and \( f' = 0 \), we can obtain the three key points of the original curve, \((t_0, H_0), (t_1, H_1), \) and \((t_2, H_2)\), where \( t_0 \) is the point at which the growth rate of the curve just starts to accelerate, \( t_2 \) is the point at which the growth rate of the curve just starts to slow down, and \( t_1 \) is the point with the maximum slope. The classification of early warning levels is shown in Table 2.

![Figure 4](image-url)
3. Experimental results and analysis

3.1. Experimental setup

Sina Weibo is a very popular application in China that is used by 300 million users monthly. In experiments, public opinion event data were collected from Sina Weibo to verify the effectiveness of the proposed scheme in terms of the correlation coefficient (R-square) and root mean square error (RMSE). Two data sets were used in the experiments, as shown in Table 3 and Table 4.

In the experiments, we consider two burst events which attract a large number of users. As shown in Tables 3 and 4, the People’s Daily review on Qiu Chen includes 102 data sets, which is on the stage of early public opinion. And Wuhan Lockdown Incident is on the stage of midterm public opinion. The numbers of forwards, comments and likes are presented as shown in Table 4. We can see the burst events get more than 500000 likes in several hours, which indicates the events vary very quickly in short term.

The following scenarios are considered in the experiments: (1) multivariate attribute prediction analysis to verify the relationship between the heat degree and multiple factors; (2) the performance comparisons of the proposed model, the approximate sparse multinomial logistic (ASML) scheme [29], and the Prophet-L scheme [30] in cases of multiple factors and small data; (3) verification of the intervention performance in the critical interval for early warning.

3.2. Multivariate attributes prediction analysis

To analyze the effects of multiple factors on public opinion, the metrics of the numbers of forwards, comments, and likes in relation to the Wuhan lockdown incident are used. The heat degree of each factor over time is shown in Figure 6.

In Figure 6, the heat degrees for the three factors of the numbers of forwards (Hot-f), comments (Hot-c), and likes (Hot-l) are presented. The maximum number of likes is 562,200, and the maximum number of comments is 23,880. The value for likes shows a marked increase over time. Therefore, it is obvious that there are large deviations among the various factors, suggesting that is unsuitable to use a single factor to predict the trend of variation.

To accurately describe this short-term bursty event, a synthesized heat degree is provided by the RVM-L model, and the results are shown in Figure 7.

As shown in Figure 7, the value of the synthesized heat degree matches the real curve very well; the fitted value floats around the real value, with an error between them of less than 6%. The reason is that the RVM maps the nonlinear multifactor relationship to a higher-dimensional space to find the internal relationships among the multiple factors. Hence, the synthesized heat degree, which conforms well to the real trend, is an effective metric for predicting the trends of variation of short-term outbreak events.

The synthesized heat degree data are then input into the logistic model for trend prediction. The predicted heat degree values obtained with multiple factors and a single factor are shown in Figure 8.

As shown in Figure 8, for the single factor-based scheme, the peak value of the heat degree is 186200, while the real public opinion peak value is 6.191 million. Thus, the value forecast using only a single factor is far lower than that of the real public opinion curve. The reason is that the overall popularity of a public opinion event is affected by multiple factors. Even though the single factor used (the number of likes) has the largest value among the various factors, it can be observed that there is still a large difference from the true public opinion value. Hence, a single factor can only partially describe the evolution of popularity in the

![Figure 5. Partial derivative results](image)

![Figure 6. Heat degree with single factor](image)

![Figure 7. Value of the synthesized heat degree](image)

![Figure 8. Predicted heat degree values](image)

3. Experimental results and analysis

| Parameter Event | Amount | Duration | Event status         |
|-----------------|--------|----------|----------------------|
| People’s Daily review on Qiu Chen | 102 sets of data | 8.5 hours | Early public opinion |
| Wuhan Lockdown Incident | 108 sets of data | 9 hours | Mid-term public opinion |

| Table 3 Data sets in the experiments |
|-------------------------------------|
| Event Parameter | Amount | Duration | Event status         |
|-----------------|--------|----------|----------------------|
| Wuhan Lockdown Incident | 108 sets of data | 9 hours | Mid-term public opinion |
| People’s Daily review on Qiu Chen | 102 sets of data | 8.5 hours | Early public opinion |

| Table 4 Descriptions on the data sets |
|-------------------------------------|
| Event Parameter | Time | The number of forwards | The number of comments | The number of likes | The value of synthesized index | The number of derived topics |
|-----------------|------|------------------------|------------------------|-------------------|-------------------------------|-----------------------------|
| Wuhan Lockdown Incident | 2021.01.23 7:00am-16:00am | 9016 | 30287 | 540731 | 6128481 | 45 |
| People’s Daily review on Qiu Chen | 2021.03.16 18:00pm-23:30am | 8132 | 20751 | 553538 | 4048247 | 17 |

![Table 2 Warning rating scale](image)
context of public opinion.

For the results obtained with synthesized multifactor heat degrees for 30 and 15 data sets, the model can predict results that are close to the real trend. The R-squares for 30 data sets and 15 data sets are 0.9973 and 0.9905, respectively, indicating that the model can still achieve high predictive performance when the amount of data is small. However, it can be seen from Figure 8 that when the peak of the predicted trend is compared with the true trend, there is still a gap between the predicted and true values.

3.3. RVM-L model performance analysis

To decrease the peak error caused by a small amount of data, the Lagrangian interpolation method is adopted to optimize the predictions of the proposed RVM-L model. In this regard, data from the Wuhan lockdown incident are used to verify the effectiveness of the proposed scheme. The error range is $V_e=10000$, and the number of terms in the interpolation polynomial is 5. The performance of the proposed RVM-L model is compared with that of the ASML scheme [29], the Prophet-L scheme [30], and the performance of the RVM-L model is also evaluated both with and without Lagrangian interpolation (lags-RVM-L and RVM-L, respectively). In addition, the point pairs (PPF) of the interpolation are presented. Interpolation is performed on the intervals [375, 400] and [400, 425].

The prediction results of the different schemes with 5 data sets and 10 data sets are shown in Figures 9 and 10, respectively, where the interpolations are performed on the intervals of [375, 400] and [400, 425], respectively. When there are only 15 sets of historical data, the corrected peak value with the lags-RVM-L model changes from 5.652 million to 5.838 million, reducing the error by approximately 3%, as seen in Figure 9. Compared to the first corrected prediction curve, the peak value in Figure 10 is closer to the true value, and its peak error is reduced by approximately 8%. Overall, the errors between the interpolated predicted values and the true values range from 1% to 6%. As shown in Figures 9 and 10, it is obvious that the interpolated predicted values are within a reasonable range of the real trend. The results also show that the proposed lags-RVM-L model can provide effective supplementary data to improve the credibility of the results. Moreover, as the amount of data increases, overfitting of the predicted values can be avoided.

The results of Figures 9 and 10 also show that the prediction effect of the RVM-L model is better than those of ASML and Prophet-L. This is due to the synthesized multifactor fitting of the RVM-L model, which reduces the error by approximately 6%. Moreover, the RVM-L results are relatively stable and conducive to trend prediction. For the ASML scheme, the fitting error is between 2% and 10%, and the fluctuations are larger when multiple factors are considered. For the Prophet-L scheme, the peak error is about 12% to 16%, the prediction ability of the model is reduced due to the insufficient data.

The RMSEs of the different schemes with 5 interpolation data sets and 10 interpolation data sets for the Wuhan lockdown event are shown in Figures 11 and 12, respectively. These results show that the RMSEs of the RVM-L and lags-RVM-L models are both lower than those of ASML and Prophet-L when public opinion is still developing and the amount of data available is insufficient, and the predicted values from the RVM-L and lags-RVM-L models are closer to the true values. For the proposed lags-RVM-L model, it is clear that the degree of dispersion between the predicted and true values is reduced by a factor of 0.5 in the case of a small amount of data.

To further verify the effectiveness of the proposed model, the prediction results of the different schemes with 5 interpolation data sets and 10 interpolation data sets for the Qiu Chen event are shown in Figures 13 and 14, respectively, where the interpolations are performed on the intervals of [135, 160] and [160, 185], respectively.

It can be seen from Figures 13 and 14 that when there are only 15 historical public opinion data sets, for the RVM-L model, the predicted trend is opposite to the real public opinion trend. However, the lags-RVM-L model can effectively forecast the trend, with errors between
the interpolated predicted values and the true values of between 1% and 6%. However, there is still an error between the predicted peak and the true peak. This is due to insufficient data collection in the early stage of public opinion development.

Figure 14 shows the results of prediction interpolation on the interval [160, 185] based on the first interpolation. The predicted trend is close to the real trend. Compared with the original predicted curve, with the first interpolation, the peak error with respect to the real trend is reduced by approximately 13%, and the error between the interpolated points in the partially enlarged graph and the real data ranges between 2% and 3%. Adding effective predicted data can improve the regularity of the historical data to enhance the model’s ability to predict future trends. After the interpolation of both intervals, the error between the predicted and true peak values is reduced to approximately 2%.

The RMSEs of the different schemes with 5 interpolation data sets and 10 interpolation data sets for the Qiu Chen event are shown in Figures 15 and 16, respectively.

Compared with the first data set in Figure 15, the second data set corresponds to the early stage data of public opinion development, and it is more difficult to predict trends when the amount of data is
insufficient. Figure 15 shows that in the initial stage of public opinion development, the RMSE of the RVM-L model is increasing, and the degree of dispersion between the predicted and true values is also increasing. In contrast, ASML and Prophet-L have lower errors between the predicted values and the true values, and the degree of dispersion markedly decreases with the proposed lagrs-RVM-L model. As shown in Figure 16, the degree of dispersion between the values predicted by each model and the true values increases over time, and compared with the values predicted by the RVM-L, ASML and Prophet-L models, the degree of dispersion of the lagrs-RVM-L results is reduced by 17.2%, 23.4% and 35.3%, respectively. These results further show the effectiveness of the interpolation method adopted in the proposed lagrs-RVM-L model.

3.4. Critical interval analysis

15 data sets from the Wuhan lockdown event are used to analyze the critical intervals for early warning. The key points found with different schemes are shown in Table 5.

As shown in Table 5, the key points vary slightly with event development, and they are all consistent with the trend changes of the original curve. Considering the nature of short-term emergencies under real circumstances, i.e., the rapid outbreak of public opinion and the small amount of data initially available, the concept of a critical interval is introduced to increase the prediction reliability for early warning regarding public opinion events. The critical intervals for this event are shown in Table 6.

From Table 6, we can obtain a low-level critical early warning interval \( t_0 \in [264, 286] \) and a high-level critical interval \( t_1 \in [305, 340] \). Under the assumption that only one intervention will be performed in each early warning interval, the effects of early warning and intervention in different intervals are shown in Figures 17 and 18. The key points of public opinion trend development with the implementation of early warning and intervention are shown in Table 7 and Table 8. The comparisons between the trend peaks with warning and intervention in different intervals and at different boundaries are calculated based on when the growth rate of the original curve begins to slow down (t is 380).

As shown in Figures 18 and 19, the trends of public opinion after warning and intervention in early warning intervals of different levels are different, and the corresponding trends for intervention at different boundaries of warning intervals of the same level are also different. With early warning and intervention at the left boundary of the low-level critical interval, the peak value is decreased by 90% compared with the original public opinion curve, and with early warning and intervention at the right boundary, the peak value is reduced by 63%. These results also demonstrate the effectiveness of the proposed early warning and intervention mechanism based on critical intervals, which can greatly reduce the hotness value of an incident. From Table 8, we can see that with early warning and intervention at the left and right interval bounds, the peak values are decreased by 52% and 16%, respectively. The results indicate that with early warning and intervention, the peak values at the three key points will decrease, the trend will be flattened, and the explosive growth of public opinion can be brought under control.

3.5. COVID-19 predictive analysis

To evaluate the performance of the proposed scheme for long-term events, a daily COVID-19-related data set from January 24 to March 30 released by the Chinese Center for Disease Control and Prevention [41] is used for analysis. The prediction results of the RVM-L, lagrs-RVM-L, ASML [29] and Prophet-L [30] models under different intervention strategies are presented.

As shown in Figure 20, the real cumulative cases showed a sharp rise on the 19th day, with the number of infections increasing from 2015 to 15,152. The occurrence of such abnormal points obviously affects the prediction accuracy of all models. Since the amount of data is very small and the infectious disease is unknown, the predicted values are considerably lower than the actual number. The errors between the actual peak value and the predicted peak values of the RVM-L and ASML models are 28% and 26%, respectively. The errors of the Prophet-L model and the proposed lagrs-RVM-L model are 3% and 7%, respectively. The Prophet-L model performed better in the prediction of the
fitting of a single variable. However, with the number of variables increasing, the ability to predict the trend will degrade. The interpolation scheme provides more accurate supplementary data to enable the model to more effectively deal with abnormal points, enhance the prediction closer to the real value.

Table 9 shows the critical intervals of the key points with the proposed lagrs-RVM-L model. According to the proposed scheme, the outbreak interval is \([16, 20]\), which means that from February 8 to February 12, the cumulative number of cases is expected to rise sharply, becoming an outbreak event. Similarly, in the interval \([40, 45]\), that is, March 3 to March 7, the number of cases is expected to gradually become stable. According to the official data, the cumulative number of confirmed cases actually increased from 44,653 to 59,804 from February 11 to February 12, corresponding to the fastest growth period of the outbreak, and the number of new infected cases remained constant starting on March 6. These results demonstrate that the predicted critical intervals are consistent with the actual case.

The performance of early warning and intervention is shown in Figure 21 and Table 10.

From Figure 21, we can see the actual outbreak curve with interventions such as quarantine, rescue, and lockdown in China. Without interventions, the trend of infected cases would have been steeper, with the number of cases increasing exponentially to a maximum value of nearly 100000. In the later stage of the outbreak, the Chinese government adopted normalized measures of prevention and control. If these measures were to be relaxed, a later rebound phenomenon could be expected, with a sharp rise to a new peak. From Table 10, it can be seen that if normalized prevention and control measures had not been adopted in the later stage, the duration of the outbreak would have been extended, with a peak value close to 150000. Therefore, China’s normalized prevention and control policy is a necessary and effective strategy.
Detailed experimental results show the effectiveness of the proposed interest or personal relationships that could have appeared to influence the range for early warning and improve the forecasting accuracy. The method of critical interval is introduced to reasonably expand the critical interval due to ‘anthropogenic emissions switch-off’ during COVID-19 lockdown in Indian cities. Sustainable Cities and Society, 62.

Kumar, Pranhabh, Hana, Sarkawt, Omidvarbonba, Hamid, Sharma, Ashish, Sahani, Jeetendra, Abhihjit, K. V., Debele, Sisay E., Zavala-Reyes, Juan C., Barwize, Yendle, & Tsivari, Aria (November 2020). Temporary reduction in fine particulate matter due to anthropogenic emissions switch-off during COVID-19 lockdown in Indian cities. Sustainable Cities and Society, 62.

El-Bawab, Tarek S. (April 2020). COVID-19 and Us. IEEE Communications Magazine, 58(4), 5.

Cerf, Vinton G. (May 21, 2020). Implications of the COVID-19 pandemic. Communications of the ACM, 63(6), 7.

Yan, Li, Zhang, Hui-Tao, Gondalves, Jorge, et al. (May 1, 2020). An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence, 2 (5), 283–288.

Hu, X., & Lu, C. (2018). Research on Key Technologies of Internet Public Opinion Monitoring and Analysis System. In 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITMEC) (pp. 165–169).

Wang, H., Wang, K., Zhao, H., & Yue, Y. (2018). Prediction of User Behavior in Smart Home Based on Improved ARIMA Model. In 2018 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 298–302).

Al-Hajla, Q. A., & Tawalbeh, L. (2019). Autoregressive Modeling and Prediction of Annual Worldwide Cybercrimes for Cloud Environments. In 2019 10th International Conference on Information and Communication Systems (ICICS) (pp. 47–51).

Yin, F., Zhang, B., Su, G., et al. (2017). Research on the public opinion pre-warnings based on the logistic model. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 381–385). IEEE.

Lu, T., Dnyaya, Z., Lixin, Y., & Pan, Z. (2015). The traffic accident hotspot prediction: Based on the logistic regression method. In 2015 International Conference on Transportation Information and Safety (ICTIS) (pp. 107–110).

Ye, Q. (Feb. 2019). A Simple Analytical Method of Gas Discharge Based on Logistic Model. IEEE Transactions on Plasma Science, 47(2), 1413–1420.

Li, D., & Liu, Z. (2015). Situation element extraction of network security based on Logistic Regression and Improved Particle Swarm Optimization. In 2013 Ninth International Conference on Natural Computation (ICNC) (pp. 569–573).

Shao, Y., & Zhang, Y. (2018). A Failure Prediction Method for Spacecraft Loads Based on Time Series Model. In 2018 12th International Conference on Reliability, Maintainability, and Safety (ICRM&S) (pp. 277–281).

Wan, J., Chen, X., Du, Y., & Jia, M. (2018). Evolutionary Prediction of Internet Propagation Dynamics Based on Time Series Characteristics. In 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS) (pp. 113–117).

Guo, Huiming, Zeng, Weiming, Shi, Yuhu, Deng, Jin, & Zhao, Le (May 2020). Kernel Granger Causality Based on Back Propagation Neural Network Fuzzy Inference System on DMRI Data. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(5), 1049–1058.

Zhang, Shunli, Zhang, Li, & Hauptmann, Alexander G. (August 2020). Fuzzy Least Squares Support Vector Machine with Adaptive Membership for Object Tracking. IEEE Transactions on Multimedia, 22(6), 1998–2011.

Morando, Alberto, Victor, Trent, & Dozza, Marco (February 2020). A Bayesian Reference Model for Visual Time-Sharing Behaviour in Manual and Automated Naturalistic Driving. IEEE Transactions on Intelligent Transportation Systems, 21(2), 803–814.

Liu, Hu-Chen, Wang, Li-En, Li, Zhuwei, & Hu, Yu-Ping (January 2019). Improving risk evaluation in FMEA with cloud model and hierarchical TOPIS method. IEEE Transactions on Fuzzy Systems, 27(1), 84–95.

Ibrahim, M N M, & Yusof, M Z M (2017). The impact of different training data set on the accuracy of sentiment classification of Naive Bayes technique. In 2017 IEEE Conference on Open Systems (ICOS) (pp. 17–20). IEEE.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1(Hun), 211–244.

An, Y., Sun, S., & Wang, S. (2017). Naive Bayes classifiers for music emotion classification based on lyrics. In IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 635–638).

4. Conclusions

To accurately predict Internet public opinion trends, an early warning scheme is proposed that comprehensively considers the propagation laws of multiple attributes of Internet public opinion and the dynamic characteristics of COVID-19-related events. A hybrid RVM-L model is proposed to predict public opinion based on multiple factors, and Lagrangian interpolation is adopted to solve the sparse data problem. The metric of critical interval is introduced to reasonably expand the range for early warning and improve the forecasting accuracy. Detailed experimental results show the effectiveness of the proposed RVM-L model for multifactor public opinion prediction as well as its ability to capture predicted trends based on sparse data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Science Foundation of China (No. 61772562), the Graduates Academic Innovation Foundation of South-Central University for Nationalities (No. 3212020syxjx131), the Key Project of Hubei Provincial Science and Technology Innovation Foundation of China (No. 2018ABB1485), and the Youth Elite Project of State Ethnic Affairs Commission of China (No. 2016-3-08).
[28] Lei, D., Zhang, H., Liu, H., Li, Z., & Wu, Y. (2019). Maximal Uncorrelated Multinomial Logistic Regression. *IEEE Access*, 7, 89924-89935.

[29] Kayabol, K. (1 Feb. 2020). Approximate Sparse Multinomial Logistic Regression for Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 490-493.

[30] Wang, P., Zheng, X., Li, J., et al. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques. *Chaos, Solitons & Fractals*, 139, Article 110058.

[31] Meng, Q. (2011). Evaluation of Residential Loan by Combining RVM and Logistic Regression. In *2011 International Conference on Management and Service Science* (pp. 1–4).

[32] Du, Y. (2018). Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network. In *2018 Chinese Control and Decision Conference (CCDC)* (pp. 2854-2857).

[33] Li, Y., & Li, S. (2013). Study on the Tibetan network public opinion dissemination behavior evolution prediction based on the cloud model. In *2013 6th International Conference on Information Management* (pp. 247-250). Innovation Management and Industrial Engineering.

[34] Kutela, Boniface, Novat, Norris, & Langa, Neema (April 2021). Exploring geographical distribution of transportation research themes related to COVID-19 using text network approach. *Sustainable Cities and Society*, 67, Article 102720.

[35] Li, Bo, Peng, You, He, He, Wang, Minghui, & Feng, Tao (March 2021). Built environment and early infection of COVID-19 in urban districts: A case study of Huangzhou. *Sustainable Cities and Society*, 66, Article 102685.

[36] Bhattacharya, Sweta, Maddikunta, Praveen Kumar Reddy, Pham, Quoc-Viet, Gadekallu, Thippa Reddy, S., Siva Rama Krishnan, Chowdhary, Chiranji Lal, Alazab, Mamoun, & Piran, Md. Jalil (February 2021). Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey. *Sustainable Cities and Society*, 65, Article 102589.

[37] Afshar, Farideh Abbasdeh, Ayoubi, Shamsollah, & Jafari, Azam (2018). The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma*, 315, 36-48.

[38] Agovino, M., Cerciello, M., Ferraro, A., & A, Garofalo (2020). Spatial analysis of wildfire incidence in the USA: the role of climatic spillovers. *Environment, Development and Sustainability*, 1–22.

[39] Bayar, Y., Sergin, H. F., Oztiirk, O. F., & Sasmaz, M.Ü. (2020). Financial literacy and financial risk tolerance of individual investors: Multinomial logistic regression approach, *SAGE Open, 10*(3), Article 2158244020945717.

[40] Cao, F., Yang, Z., Ren, J., Ling, W. K., Zhao, H., & Marshall, S. (2017). Extreme sparse multinomial logistic regression: A fast and robust framework for hyperspectral image classification. *Remote Sensing*, 9(12), 1255.

[41] Source of epidemic data: [http://www.chinacdc.cn/](http://www.chinacdc.cn/), 2020.