

A Study on Predicting Social Media Opinion Dynamics Based on Time Series Data Modeling Techniques

Cheng Tang^{1,*}

¹ School of Humanities, Hunan City University, Yiyang, Hunan, 413000, China

Corresponding authors: (e-mail: 18878970867@163.com).

Abstract With the rapid development of social media, public opinion information on social platforms has shown explosive growth. Accurately predicting the trend of social media public opinion dynamics is of great significance in grasping the direction of public opinion development and intervening in the dissemination of undesirable public opinion in a timely manner. This study explores the prediction method of social media opinion dynamics based on time series data modeling technology. First, a multilevel prediction model integrating theme identification layer, feature processing layer, prediction layer and parameter optimization layer is constructed. The LDA model is used to identify the themes of social media public opinion, and feature splicing is used to complete the fusion of multi-thematic features. In the prediction layer, the multidimensional features are used as input variables, the LSTM model is used to realize the dynamic prediction of public opinion, and the model hyperparameters are optimized by the gray wolf optimization algorithm. The experimental results show that the correlation coefficient of the optimized LSTM model in this paper reaches 0.518 on the public opinion dataset, which is significantly higher than that of the comparison models such as ARMA, Prophet, Informer and the original LSTM, and the standard error RMSE is 2012.117, and the average absolute percentage error MAPE is only 4.275, which is about 30% lower than that of the comparison models. In the prediction of information dissemination in “a hot spot”, the MSLE value of this model is reduced by 0.124 and the MAPE value is reduced by 0.056 compared with the optimal comparative model Informer, and the study shows that the temporal data modeling method integrating multi-topic features can effectively improve the accuracy of the prediction of the dynamics of the public opinion in social media, and it has practical application value for the early warning and intervention of public opinion changes in the hot spot events in the society.

Index Terms social media, opinion dynamics, LSTM model, gray wolf optimization algorithm, time-series data, multi-topic feature fusion

1. Introduction

With the continuous expansion of the scale of Internet users, the Internet has become another important way of information dissemination in addition to traditional media such as newspapers, radio and television. At present, social media such as microblogging, Xiaohongshu, Facebook, Twitter and other social media with the help of the Internet platform have become an important battlefield for information dissemination. When unexpected online events or social events attract a lot of attention from netizens and evolve into online public opinion, especially negative online public opinion, this will have a significant impact on social public safety and long-term development [1]-[4]. By analyzing the trend of public opinion on specific topics, identifying the key nodes of public opinion dissemination, and modeling and predicting the future direction of public opinion, we can provide important support for grasping the overall trend of public opinion and recognizing the public opinion environment [5]-[7]. In the era of big data, public opinion information work needs to collect a huge amount of data. At the same time, social media and other network platforms make network communication present a “honeycomb” dispersive structure, which makes it more difficult to analyze and predict network public opinion, and the warning time is unstable with the outbreak of public opinion events, which makes the response to network public opinion face greater challenges [8]-[11]. In addition, public opinion trend prediction usually starts from public opinion analysis to predict future changes in netizens' opinion positions. However, due to the large number of active users in social media, the large amount of content data involved, as well as the many factors such as the heterogeneity of the social network itself and the variability of public opinion fermentation, the prediction results are often not very accurate [12]. While the social media public opinion's is an ever-changing phenomenon, by analyzing the time-series data in social media and

using the time-series data modeling technology to construct and predict its model to reveal the law and trend of public opinion and get better prediction results [13].

In the prediction of public opinion, the monitoring of online public opinion is mainly realized through techniques such as topic detection analysis and content-based sentiment classification analysis. Literature [14] emphasized the sentiment vectors in microblogging emergencies, performed sentiment analysis with fuzzy neural networks, and constructed a hybrid public opinion prediction model by combining the gray prediction model. Similarly, literature [15] analyzed and classified the textual sentiment in microblog posts, and used the maximum entropy model and the principle of minority to majority for microblog opinion prediction, obtaining 88% prediction accuracy. Literature [16] quantifies user sentiment changes in online public opinion using the damped oscillator model and simulates them through examples, and extracts key indicators for public opinion events with the help of particle swarm optimization algorithm, so as to predict the dynamic changes of public opinion sentiment.

In addition, literature [17] streamlined the monitoring parameters of network public opinion under rough set theory, and formulated a new network public opinion monitoring system based on reality, combined with the hierarchical analysis method to determine the weights of the parameters, and predicted the network public opinion from a quantitative and qualitative perspective with the fuzzy comprehensive evaluation model. Literature [18] designed a seasonal grey decomposition and integration model for online public opinion prediction, which was optimized by STL decomposition algorithm, grey difference information, dynamic seasonal factors and Bernoulli's equation to obtain higher prediction accuracy. Literature [19] designed a multi-node network by putting users in the form of nodes, tracked and categorized user opinions and used events and opinions in social media to generate opinion hit matrices to obtain real-time opinion prediction in social media. Literature [20] referred to the susceptibility-infection-recovery model and constructed an opinion dynamics model for link prediction, which upgraded the public opinion prediction and link prediction accuracy, and effectively managed the online opinion.

Literature [21] proposed an online public opinion prediction model supported by the gray prediction model, which mediates the addition of public opinion fluctuations, upgrades the model performance, and further improves the prediction accuracy of the model. Literature [22] created a hybrid model with an enhanced particle swarm optimization algorithm and optimized long and short-term memory networks to predict the trend of public opinion dissemination in emergencies, which improved the performance of the model by 13.59%-74.27% over the base model. Literature [23] used the vulture algorithm to improve the parameters in the radial basis function neural network, and in this way contributed a model for predicting the trend of online public opinion, which has better accuracy and stability. Literature [24] introduced supervised machine learning and deep learning to categorize the political sentiments of users in Twitter as a way to predict the online public opinion of political parties in Pakistan. Literature [25] utilized convolutional neural network, bi-directional long and short term memory network, SoftMax classifier for local and global feature extraction of social media opinion texts and in this way achieved 95.84%-97.56% accuracy in predicting the dynamics of public opinion sentiment.

Social media has become an important platform for people to express their views and exchange ideas, as well as an important place for the formation and dissemination of public opinion. With the expanding scale of users and the continuous increase in the frequency of use, the public opinion information on social media has shown explosive growth, and these massive public opinion data have far-reaching impacts on government departments, corporate organizations and individuals. Accurately grasping the dynamic trends of social media public opinion is important for early warning of potential public opinion risks and formulating effective response strategies. However, social media public opinion is characterized by fast dissemination speed, wide influence range, and complex and variable content, which makes the prediction of public opinion dynamics face many challenges. Traditional public opinion analysis methods often rely on static data, which is difficult to capture the dynamic characteristics of the evolution of public opinion; it is also difficult to comprehensively reflect the diversity and complexity of public opinion by analyzing a single feature dimension. Therefore, how to build a high-precision prediction model of social media public opinion dynamics based on time-series data modeling technology, taking into account multi-dimensional features, has become a key issue in current research. In recent years, deep learning techniques have made significant progress in the field of time-series data analysis, especially Long Short-Term Memory Network (LSTM) has been widely used in time-series prediction tasks due to its ability to effectively deal with long sequence dependency problems. Meanwhile, social media opinion often involves multiple topics, and there are mutual influences and interactions among the topics, so how to effectively fuse the multi-topic features to improve the prediction accuracy is also a direction worth exploring. In addition, the selection of model parameters has a crucial impact on the prediction performance, and the Gray Wolf Optimization (GWO) algorithm, as a new type of intelligent optimization algorithm, shows good performance in parameter optimization. Based on the above analysis, this study combines the time-series data modeling technique with topic identification, feature fusion, and parameter optimization to construct a multi-level dynamic prediction model for social media opinion. First, the LDA topic model

is used to identify the topics of social media data and obtain the topic-vocabulary distribution and topic-text distribution of social hotspot events; second, each topic feature is processed according to the quantization methods of different features, and the splicing method is used to realize the multi-topic feature fusion; once again, the multidimensional features are used as inputs, and the number of original microblogs is used as the output sequence, to construct the LSTM-based public opinion dynamics. Finally, the hyperparameters of the LSTM model are optimized using the Gray Wolf optimization algorithm to improve the prediction accuracy. Finally, the hyperparameters of LSTM model are optimized by the Gray Wolf optimization algorithm to improve the prediction accuracy. The prediction effects of different algorithms are compared and analyzed through experiments to verify the effectiveness and superiority of this model in predicting the changes of opinion dynamics in social media.

II. Core methodology

II. A. LSTM model structure

The first step in the execution step of the LSTM network [26], [27] is to characterize the specific data to determine which irrelevant information from the previous cell state should be most forgotten at the current time. The forgetting gate f_t outputs a vector with each element belonging to the interval (0,1) by inputting the information from X_t and h_{t-1} , indicating the degree of retention that needs to be preserved in the cell state C_{t-1} . b_f and W_f are the bias term and input weight of the forgetting gate f_t , respectively. The specific expression for f_t is then:

$$f_t = \sigma(W_f \cdot (X_t, h_{t-1}) + b_f) \quad (1)$$

The second step is to determine which input information should best match the current cell state by specific information changes. First, X_t and h_{t-1} in the input gate i_t in σ to determine the specific change in input information. Then X_t and h_{t-1} are given new candidate cells \tilde{C}_t by the activation function \tanh , which produce transformations on the cell's information update. The specific computation is as follows:

$$\begin{aligned} i_t &= \sigma(W_i \cdot (X_t, h_{t-1}) + b_i) \\ \tilde{C}_t &= \tanh(W_c \cdot (X_t, h_{t-1}) + b_c) \end{aligned} \quad (2)$$

The principle of their operation can be summarized as a bitwise multiplication operation based on the tanh function, which can be expressed by the following equation:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3)$$

The tanh function, through its own nonlinear characteristics, can better fit the nonlinear model and improve the expressive ability of the neural network. And its value domain is $[-1, 1]$, which can limit the output value to a certain range and avoid the problem of gradient explosion. Therefore, tanh is chosen as the activation function to expect to achieve better model performance.

The third step is to update the current state of the memory unit. By calculating the dot product of the forgetting gate f_t and the memory unit C_{t-1} of the previous moment, this process ensures that only the relevant historical information is retained in the memory unit, which provides the basis for subsequent information processing. Meanwhile, the new candidate cell information to be added is determined based on the dot product of the candidate cell state \tilde{C}_t and the input gate i_t , which is calculated in the second step. C_t is calculated as follows:

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (4)$$

The final step is to compute the value of the output gate O_t after obtaining the first-order updated value of the cell state from the above computation. Where b_o and w_o are the bias term and input weight of the output gate O_t , respectively, and the final output is determined by O_t and C_t . The computational formula is:

$$\begin{aligned} O_t &= \sigma(W_o \cdot (X_t, h_{t-1}) + b_o) \\ h_t &= O_t * \tanh(C_t) \end{aligned} \quad (5)$$

II. B. Gray Wolf Optimization Algorithm

The Gray Wolf Algorithm (GWO) [28], [29] is a new type of intelligent optimization algorithm inspired by the process of rounding up and hunting prey from the gray wolf population, and the leader of the wolf pack is called alpha wolf, abbreviated as α . The α wolf is mainly responsible for making decisions about hunting and directing the activities of the pack, and plays a dominant role over the entire wolf pack. The second level of the wolf hierarchy is the assistant wolf beta, abbreviated as β . The β wolf is responsible for providing some decision-making and

assistance to the α wolf during the hunting process, and the permanent mentor is responsible for guiding the leader of the pack. The β wolf reinforces the α wolf's commands in the pack and provides feedback to the α wolf.

The lowest ranking wolf in the pack is the base wolf omega, abbreviated as ω . The ω wolf is often seen as a scapegoat and is forced to bow down and submit to all dominant wolves. If a wolf is not α , β , or ω , it is therefore named the subordinate wolf delta, abbreviated as δ . The δ wolves must obey α and β , but they dominate ω . Help α and β in hunting prey and providing food for the pack.

As in Fig. 1, the main idea of the GWO algorithm is as follows: starting from any position in the space to be searched for superiority, the individual with the optimal fitness is named as the head wolf α , the individuals with slightly lower fitness in the 2nd and 3rd are named as helper wolves β and subordinate wolves δ , and the rest are the base wolves ω . In the process of rounding up prey by the gray wolf population, β and δ assist α in judging the location of the prey, and when the location of the prey is found, α , β , and δ lead ω in rounding up the prey, and outputting the optimal value through an iterative search time and time again.

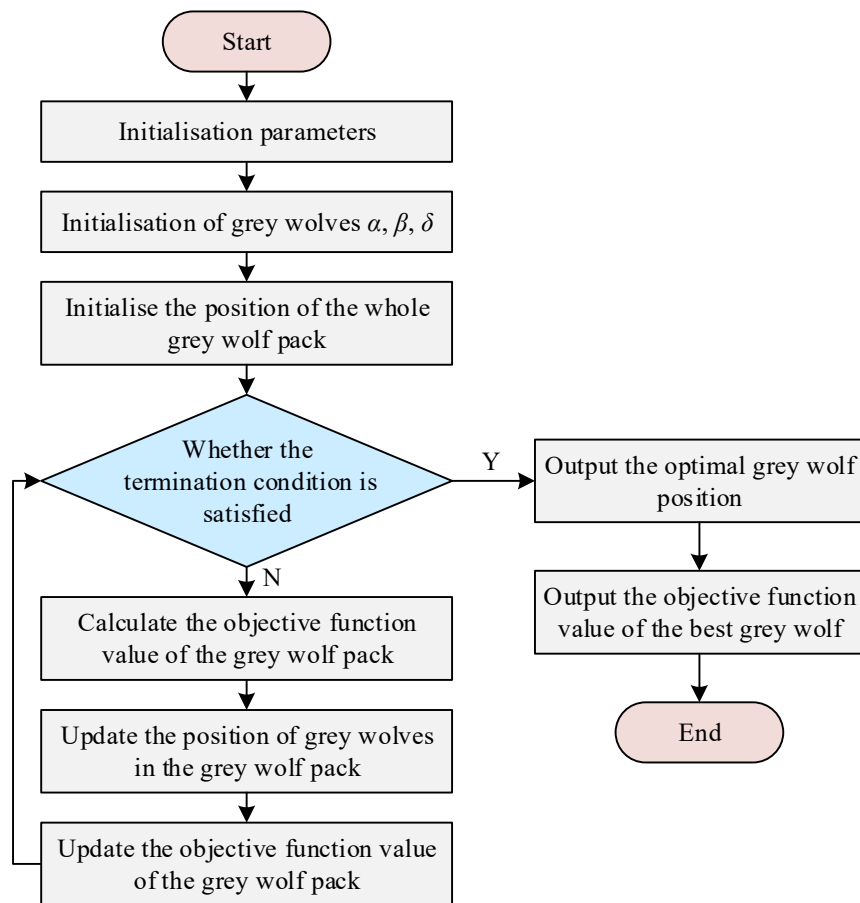


Figure 1: GWO algorithm flowchart

III. Predictive modeling of public opinion dynamics in social media

The dynamic prediction model of social media public opinion fused with multi-topic features constructed in this paper is shown in Figure 2. It consists of four layers: theme identification layer, feature processing layer, prediction layer, and parameter optimization layer, and each layer progresses layer by layer, obtaining the online public opinion themes of social hotspot events based on the theme model, completing the fusion of multi-topic features based on the topic-associated perspective, predicting the dynamics of public opinion based on the multivariate time series analysis, and selecting the optimal parameters of the prediction model based on the optimization algorithm. In the theme recognition layer, the theme recognition method of online public opinion on hot social events is used to discover the issues of public concern in hot social events, and to obtain the theme-vocabulary distribution and

theme-text distribution of hot social events. In the feature processing layer, different features are processed according to their quantization methods, and then the features of each theme are fused and spliced to obtain multi-topic fusion features. In the heat prediction layer, content attention, sentiment score, number of followers, number of fans, opinion leader participation, theme weight, and theme-word importance are used as relevant factors, and the number of original microblogs is used as the output sequence, and the dynamic prediction of online public opinion on social hotspot events is realized using the LSTM model. In the parameter optimization layer, the GWO algorithm is used to iteratively optimize the hyperparameters such as the number of neurons, the number of iterations, and the initial learning rate of the prediction model in order to further improve the accuracy and precision of the dynamic prediction results.

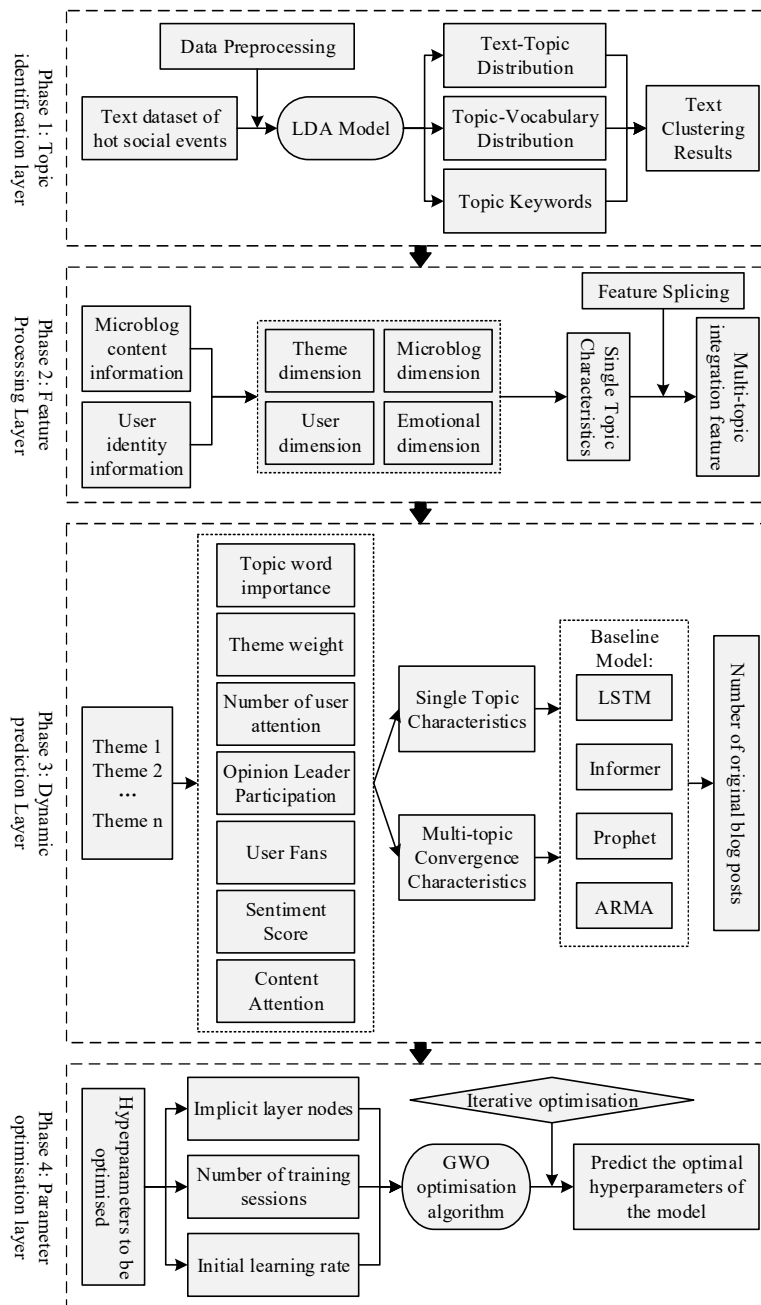


Figure 2: Network public opinion dynamic prediction model

III. A. Social Media Opinion Theme Identification

In this paper, the LDA model is used to complete the work of identifying the theme of online public opinion on social hotspot events, and the specific process can be divided into the following four steps:

In the first step, for the j th topic, the distributional probability vector $\varphi^{z_j} = Dir(\beta)$ for the polynomials of the feature words on the topic is computed via the Dirichlet distribution, and N is selected using the Poisson distribution $N = Poisson(\xi)$, where N represents the length of a single text.

In the second step, θ is determined based on $\theta \sim Dir(\alpha)$, where θ obeys the *Dirichlet*(α) distribution, θ denotes the probability of the occurrence of each topic, and α is a parameter of the *Dirichlet* distribution.

In the third step, a topic z_n is selected for each of the k feature words w contained in each text, z_n obeys a *Multinomial*(θ) multinomial distribution over the topic distribution vector θ , and z_n is the randomly selected topic. A feature word w_n is randomly selected from the above selected topic z_n according to $p(w_n | z_n; \beta)$.

p is the multinomial distribution of the topic z_n . β is a $K \times N$ dimensional matrix. $\beta_{ij} = P(w_i = 1 | z_j = 1)$, indicating that β is the probability of generating the feature word w_i for the record topic z_j .

The generation probability $p(w_i)$ of w_i in the hot event text d is:

$$p(w_i) = \sum_{j=1}^T P(w_i | z_i = j) P(z_i = j) \quad (6)$$

where w_i represents the i th topic that appears in topic z_i . $p(w_i | z_i = j)$ represents the probability of occurrence of feature word w_i in topic z_i . $p(z_i = j)$ is the probability that topic z_i appears in the document. Therefore, the probability $p(w | d)$ that feature word w appears in text d is:

$$p(w | d) = \sum_{j=1}^T \varphi_w^j * \theta_j^d \quad (7)$$

In the fourth step, the EM algorithm is used to obtain the approximate solution of α , β , which in turn builds the LDA three-layer model:

$$l(\alpha | \beta) = \sum_{i=1}^M \log P(d_i | \alpha, \beta) \quad (8)$$

where $P(d | \alpha, \beta)$ denotes the conditional probability distribution of the generated document d , which is calculated as follows:

$$P(d | \alpha, \beta) = \frac{X}{Y} \int \left(\prod_{i=1}^T \theta_i^{\alpha_i - 1} \right) \left(\sum_{n=1}^N \sum_{i=1}^T \sum_{j=1}^M (\theta_j \beta_{ij})^{w_{ij}^d} \right) d\theta \quad (9)$$

The LDA model is constructed through the above process to complete the topic recognition task and output the text-topic matrix and topic-topic word matrix.

III. B. Convergence of Multi-topic Characteristics of Public Opinion on Hot Spot Events

In this paper, when predicting online public opinion dynamics for a topic in a hot social event, we will add the features of other topics under the event into the input variables of the prediction model, and compare and analyze the differences in the prediction results when the single topic features and multi-topic fusion features are used as the input variables respectively, in order to explore the role of expanding the feature dimensions of the input variables in improving the accuracy of the prediction model.

In this paper, we adopt the splicing method to fuse the features of multiple themes with multiple dimensions in hot social events:

$$Merge = \sum x_i^j \quad (10)$$

where i represents each topic in the event, j represents the feature corresponding to each topic, and x_i^j denotes the j th feature under the i th topic.

III. C. Evaluation of Opinion Dynamics Prediction Models

In order to measure the fitting effect of the dynamic prediction model of online public opinion as well as to facilitate the quantitative comparison of the prediction results with real data, the model evaluation in this paper is divided into two parts. The first part is the prediction algorithm evaluation: comparing the prediction results of the baseline models ARMA, Prophet, Informer, and LSTM dynamics under the input scenario of a single topic feature variable, and evaluating the effectiveness of the optimized LSTM model in dynamic prediction in this paper. The second part is input variable evaluation: single topic features and multi-topic fusion features are respectively used as input variables for the optimal prediction algorithm, and the optimization effect of introducing other topic features under the same event on the dynamic prediction model is analyzed through comparative experiments.

III. D. Dynamic prediction model optimization process

In this paper, we choose to use the Gray Wolf optimization algorithm, which has very strong local search capability and fast convergence speed, to iteratively optimize the hyperparameters of the LSTM model. The specific steps are as follows:

In the first step, the hyperparameters to be optimized by the LSTM algorithm are determined, which are the hidden layer nodes, the number of training times, and the initial learning rate. Then, determine the four initial parameters of the gray wolf optimization algorithm: search space, maximum number of iterations, dimension t_{\max} , dimensionality D , and the gray wolf population size N , the search space, i.e., the ranges of values of all the hyperparameters. Finally, the position vector x_i of each gray wolf in the population is initialized. Let the current population be $x_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}^j$, x_{ij} denotes the spatial dimension where the i th individual in the population is located, and $j = 1, 2, \dots, D$.

In the second step, the three parameters to be optimized by LSTM correspond to the vectors of the location of each gray wolf in the population, and the fitness values of the location vectors are calculated according to Eq. (11).

$$f = \sum_{i=1}^n (y'_i - y_i)^2 \quad (11)$$

In the third step, the gray wolf position vectors corresponding to the three smallest fitness values in the population are filtered and assigned to $x_\alpha(t)$, $x_\beta(t)$, $x_\delta(t)$, and by updating the $x_\alpha(t)$, $x_\beta(t)$, $x_\delta(t)$ of the position vectors so that the fitness is always minimized. When the maximum number of iterations is reached, the update is stopped. Instead, the search for the optimal value continues.

In the fourth step, after the iteration, the hidden layer nodes, the number of training times, and the initial learning rate of the LSTM neural network are updated to the values of the optimal x_α position vectors, respectively, and then the model is trained and predicted according to the new parameters.

The GWO optimization algorithm is used to find the optimal initial parameters of the LSTM, with the goal of minimizing the root mean square error RMSE of the test set.

IV. Model optimization and experimental comparison

IV. A. Experimental environment

The model establishment and operational testing configuration of this experiment is as follows: Processor is Intel(R) Core(TM) i5-3210M CPU Dual-core @2.50GHz 2.50. 4.00GB RAM. 500GB hard drive. Windows 10 64-bit operating system. Development platforms are MATLAB R2016b and R3.3.3.

IV. B. Data sources and pre-processing

IV. B. 1) Data sources

Sina Weibo is the largest anonymous social platform in China, which is public and has media attributes. With its fast spreading speed and high influence, it is the origin and spreading gathering place of most public opinion events in China. Sina Weibo's micro-index is a series of indicator products reflecting the development status of different event areas based on Weibo user behavior data, massive blog post data, and scientific algorithms, and the product mainly consists of two major modules: the Hotword Index and the Influence Index. In this paper, we adopt the Hotword Index as the quantitative value of public opinion trend, which is referred to as Micro Index. Sina Weibo's Hotword Index is an indicator that scientifically analyzes and calculates the long-term trend of the weighted sum of the frequency of mentions of each keyword in blog posts on the microblogging platform based on the daily tweets of a large number of microblogging users, and presents it in the form of a curve graph. The micro-index represents the degree of buzz about the keywords of public opinion among netizens on the microblogging platform, so the micro-index represents the degree of buzz about the public opinion events on microblogging.

IV. B. 2) Data pre-processing

For time series data, the first step is to carry out the preprocessing process of the sequence, i.e., to determine whether the sequence data is a purely random sequence and whether the data is smooth. Pure random sequence, i.e., it is a randomly generated sequence without any regularity, which is also known as white noise sequence. It is a smooth sequence with no information to be extracted, and there is no need for sequence analysis, let alone modeling prediction.

The time series of public opinion events are not white noise series, and then analyze whether the series are smooth or not. In this paper, the collected data were subjected to Augmented Dickey-Fuller test [30] in MATLAB tool, and by calculating the unit root of each index, the test results showed that the sequence data of the two public

opinion events were non-smooth sequences. Normalization of public opinion data. The normalization is done by restricting the values of the time series data to the interval $[-1,1]$, and for the time series $x_i (i = 1, 2, \dots, n)$ the min-max normalization is applied as in Eq:

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{12}$$

where x_{\max} and x_{\min} are the maximum and minimum values in the opinion data, respectively, and x'_i is the i th time series value after normalization.

IV. C. Predictive analysis of public opinion dynamics based on time series data

IV. C. 1) Forecasts of trends in the dynamics of user viewpoint positions

User position detection is performed based on the acquired basic data, and position trend change analysis is carried out. The user position detection in this paper is mainly for the current topic, detecting the user position of post commenters, and the position is divided into three attitudes of support, opposition and neutrality to the topic.

After analyzing the stance of individual comments, it is necessary to count the overall percentage of the number of people with stance, and realize the prediction of the trend change of the stance percentage through the constructed dynamic prediction model of public opinion. Position ratio is the ratio of the number of posts attributed to a certain position to the total number of posts in a certain period of time for analyzing the position trend of all comments under the topic. The predicted results of the trend change in the number of people in each stance as a function of the total number of posts are shown in Figure 3. From the figure, it can be seen that with the increase in the number of posts, the ratio of the number of people who support, oppose, and neutral attitudes converges to 0.53, 0.33, and 0.14, respectively. The implementation logic of the model trend change prediction is that the comment posts supported by each stance analyze the stance and then add them to the set of each stance. When a new user comments, according to the views held by the corresponding set. When a user who has already spoken speaks again, the set remains unchanged if the viewpoints held remain unchanged, and if the viewpoints held change, the person is deleted from the original set and added to the set of people whose new positions are held.

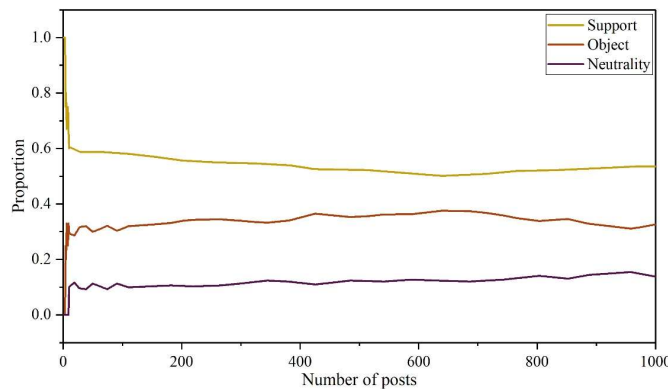


Figure 3: The number of positions is the result of the trend change of the total post

IV. C. 2) Dynamic analysis of changes in public opinion influence

This section uses the opinion dynamics model constructed above for social media opinion influence prediction. The model takes into account both the functional relationship between a time point and the previous time point in the time series data, as well as the error part of the function fitting.

The model's fitting and prediction results for social media opinion data are shown in Figure 4. In the figure, the brown fold is the effect of the model fit, and the yellow fold is the real opinion data time series plot. The separate brown part is the prediction result of the model. From the figure, it can be seen that the model's fitted folds on the original data have a high degree of overlap with the real folds, which can accurately predict the dynamic trend of the heat of the opinion data. Comparing the two folds, it can be seen that the prediction of the influence of public opinion data on the more influential users of public opinion data can accurately predict the fluctuation of the heat of public opinion at each point in time, and the prediction results of the stronger heat of public opinion at the point in time also have a more obvious upward trend.

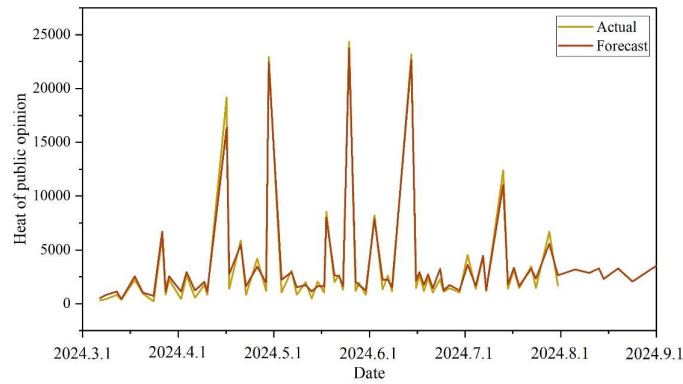


Figure 4: The fitting and prediction effect of social media public opinion data

This section statistically analyzes and evaluates the prediction results by using Correlation Coefficient (Corr), Standard Error (RMSE), and Mean Absolute Percentage Error (MAPE) as evaluation metrics. Influence prediction experiments were conducted on the opinion data collected above using Autoregressive Integral Sliding Average (ARMA) model, Prophet model, Informer model, and original Long Short-Term Memory Network (LSTM) model with the model of this paper, and two sets of experimental data were obtained. The Corr, RMSE, and MAPE values are calculated for the prediction data, and the running time of the model is counted during the model fitting and prediction process, and the statistical results of the evaluation indexes of model fitting are obtained as shown in Table 1.

By analyzing the Corr in the statistical results it can be seen that the fitting result of this paper's model is 0.518, and the similarity in the dataset is significantly higher than the comparison model. From the statistical data, it can be seen that the overall RMSE is higher, but the prediction of public opinion is essentially an estimation of the development of public opinion, and the focus is on observing whether the predicted trend at each point in time matches. Observing the RMSE of different models, it can be seen that the fitting error of this paper's model is the lowest on the dataset, with an RMSE of 2012.117, and the smaller the RMSE is, the better the model fits. MAPE is a measure of the deviation between the predicted value and the real value. As can be seen from the statistical results, the MAPE of this paper's method on the dataset is 4.275, which is much lower than that of the comparison model, which indicates that the model can capture the opinion hotspots on the data very well.

Table 1: Model fitting evaluation index

Model	Corr	RMSE	MAPE	Time (h)
ARMA	0.327	2783.179	7.341	5.31
Prophet	0.413	2413.265	6.442	4.25
Informer	0.376	2962.173	7.176	4.18
LSTM	0.339	2514.436	8.243	5.44
Ours	0.518	2012.117	4.275	2.53

IV. C. 3) Forecasting the dynamic dissemination of information in the public opinion arena

In this subsection, the model is used to predict the dynamic propagation of public opinion messages using the data of “a hot spot” introduced in the public opinion dataset collected above. The real opinion dataset is used to further demonstrate the significance and validity of the model, and the results of the model prediction are analyzed at the same time.

In this paper, a total of 217,953 original microblogs were captured in the opinion forum of “a hot spot”, and the average number of retweets per microblog was 7.57. The sum of the number of times all messages were retweeted during the observable time, i.e., from March 6, 2024 to September 22, 2044, was 2791073. The total number of tweets with 0 retweets was 72,451. In the training of the public opinion dynamic prediction model, the microblogs with 0 retweets are deleted, and the empirical data of the public opinion field of “a hot spot” are shown in Table 2.

(1) Cascade data analysis

This subsection analyzes the cascade data distribution of “a hotspot” presented in the table above. Figure 5 shows the results of the cascade distribution of “a hot spot”. From the figure, it can be seen that the cascade distribution has serious unevenness in the public opinion field. That is, most of the microblogs are retweeted less often, and only a very few microblogs are widely spread. Therefore, it can be assumed that in the cascade graph of information dissemination in the public opinion field, a few nodes have a very large number of connected edges,

and most nodes have only a small number of edges. That is, the opinion field information propagation cascade graph is characterized by scale-free network.

Table 2: "A hot spot" public opinion field empirical data

Statistical term	Statistical value
Node number	121864
Cascade number	36063
Side number	2060966
Average cascade length	57149

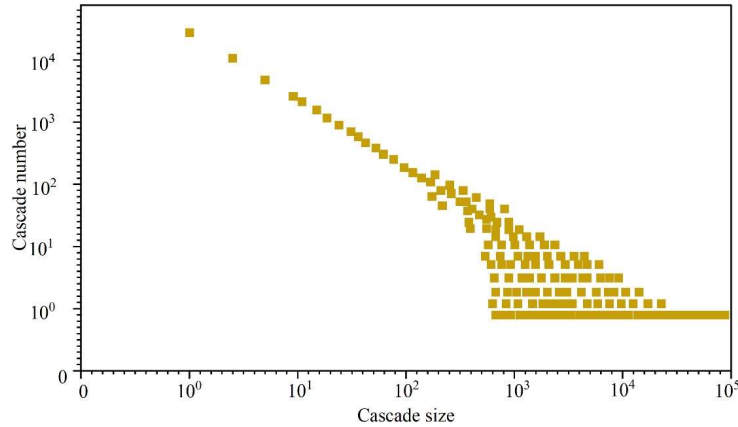


Figure 5: The "A hot spot " cascade distribution results

(2) Propagation scale analysis

In this paper, we first validate the performance of the model from the dataset presented in the above table as input data. In addition, the trained model is then used to predict the scale of information dissemination in the public opinion forum in October, and its prediction results are analyzed to further observe the trend of the public opinion forum.

In order to prove the effectiveness and superiority of this paper's model in predicting information dissemination in real large-scale public opinion forums, it is compared with several existing comparative methods, including: four comparative methods: ARMA, Prophet, Informer, and LSTM. The predicted MSLE and MAPE for "a hotspot" information dissemination are shown in Figure 6. The results show that the predicted results of the model proposed in this paper on MSLE and MAPE are smaller than the other comparison methods in predicting the scale of information dissemination. The Prophet model with the worst model performance predicts results with MSLE and MAPE values that are 0.354 and 0.103 larger than those of this paper, respectively. Among the four compared contrasting methods, the Informer model performs well, and comparing this paper's model with it, the MSLE value decreases by 0.124, while the MAPE value decreases by 0.056. Taken together, the results demonstrate the effectiveness of this paper's model for information dissemination in the public opinion arena.

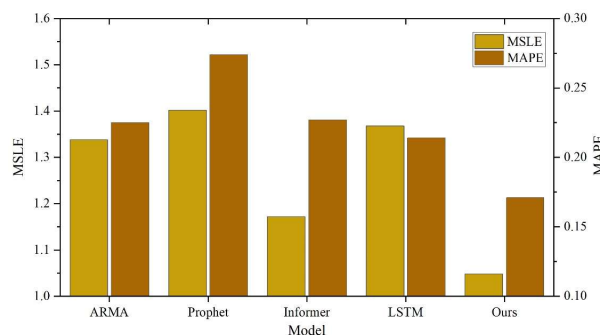


Figure 6: "A hot spot" information dissemination forecast MSLE and MAPE

In addition, the results of this paper's analysis of the predicted scale of all nodes compared with the current scale are shown in Fig. 7. Where the horizontal coordinate in the figure indicates the current scale of node propagation, the vertical coordinate indicates the predicted scale of this paper's model, the straight line is $y=x$, and each point indicates a node. The closer a node is to the straight line, the less often that piece of information will be forwarded in the future. It can be found that all nodes are almost close to the straight line, i.e., the message published in October will not be massively spread again. It is also found that it can also be noticed from the graph that the smaller the current size of the nodes, the closer they are to the straight line, i.e., the less likely these nodes are to be forwarded again in the future.

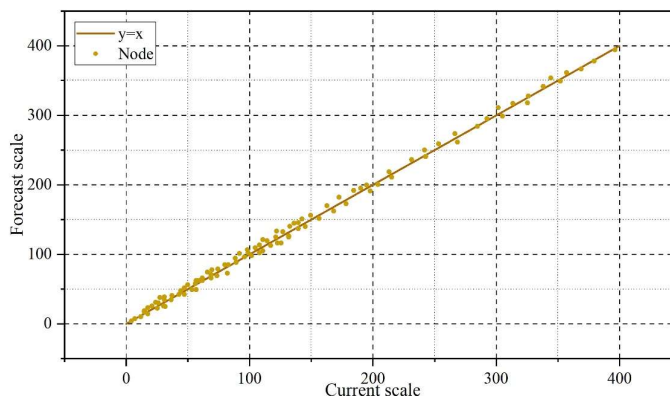


Figure 7: The prediction of all nodes and the current scale

V. Conclusion

In this study, a social media opinion dynamics prediction model fusing multi-topic features is constructed based on time-series data modeling techniques, and accurate prediction of opinion dynamics is achieved through four layers: topic identification, feature fusion, LSTM prediction and GWO parameter optimization. The results show that the optimized LSTM model in this paper fits significantly better than the traditional method on the experimental dataset, and the correlation coefficient reaches 0.518, which is 0.105 higher than the closest Prophet model; In terms of operational efficiency, this model takes only 2.53 hours to complete the modeling and prediction, which saves 52.4% and 53.5% of time compared to ARMA and LSTM models, respectively. In the prediction of information dissemination scale of “a hot spot”, the MSLE and MAPE of this model are better than the comparison model, and compared with the best performance Informer model, the MSLE and MAPE are reduced by 27.6% and 25.1% respectively. In addition, by predicting the dynamic trend of users' opinion stance, it is found that with the increase of posting volume, the proportion of the number of people with the three attitudes of supportive, opposing, and neutral converges to 0.53, 0.33, and 0.14 respectively, which provides a quantitative basis for grasping the direction of public opinion. The study proves that the time-series data modeling method integrating multi-topic features can effectively improve the accuracy and efficiency of predicting the dynamics of social media public opinion, which is of great theoretical value and practical application for understanding the law of public opinion dissemination, early warning of the risk of public opinion, and formulating intervention strategies. Future research can further explore finer-grained feature extraction methods and model adaptability in more complex scenarios to cope with the increasingly diverse social media public opinion environment.

References

- [1] Chen, Y., Li, Y., Wang, Z., Quintero, A. J., Yang, C., & Ji, W. (2022). Rapid perception of public opinion in emergency events through social media. *Natural hazards review*, 23(2), 04021066.
- [2] Gabore, S. M., & Xiujun, D. (2018). Opinion formation in social media: The influence of online news dissemination on Facebook posts. *Communicatio*, 44(2), 20-40.
- [3] Chu, M., Li, H., Lin, S., Cai, X., Li, X., Chen, S. H., ... & Chiang, Y. C. (2021). Appropriate strategies for reducing the negative impact of online reports of suicide and public opinion from social media in China. *Frontiers in public health*, 9, 756360.
- [4] Zhou, Z., Zhou, X., Chen, Y., & Qi, H. (2024). Evolution of online public opinions on major accidents: Implications for post-accident response based on social media network. *Expert Systems with Applications*, 235, 121307.
- [5] Baum, M. A., & Potter, P. B. (2019). Media, public opinion, and foreign policy in the age of social media. *The Journal of Politics*, 81(2), 747-756.
- [6] Brooks, H. (2024). Monkeying around on Twitter (now X): How big data and social media can be harnessed by zoos and other animal care facilities to examine public opinion trends. *Journal of Zoo and Aquarium Research*, 12(3), 185-194.

- [7] Yang, Y., Fan, C., Gong, Y., Yeoh, W., & Li, Y. (2024). Forwarding in Social Media: Forecasting Popularity of Public Opinion With Deep Learning. *IEEE Transactions on Computational Social Systems*.
- [8] Wang, X., Xing, Y., Wei, Y., Zheng, Q., & Xing, G. (2020). Public opinion information dissemination in mobile social networks—taking Sina Weibo as an example. *Information Discovery and Delivery*, 48(4), 213-224.
- [9] Safarnejad, L., Xu, Q., Ge, Y., Krishnan, S., Bagarvathi, A., & Chen, S. (2020). Contrasting misinformation and real-information dissemination network structures on social media during a health emergency. *American journal of public health*, 110(S3), S340-S347.
- [10] Vepsäläinen, T., Li, H., & Suomi, R. (2017). Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections. *Government Information Quarterly*, 34(3), 524-532.
- [11] Fang, S., Zhao, N., Chen, N., Xiong, F., & Yi, Y. (2019). Analyzing and predicting network public opinion evolution based on group persuasion force of populism. *Physica A: Statistical Mechanics and Its Applications*, 525, 809-824.
- [12] Wang, Z., Zhang, S., Zhao, Y., Chen, C., & Dong, X. (2023). Risk prediction and credibility detection of network public opinion using blockchain technology. *Technological Forecasting and Social Change*, 187, 122177.
- [13] Somyanonthanakul, R., Warin, K., Amasiri, W., Mairiang, K., Mingmalairak, C., Panichkitkosolkul, W., ... & Suebnukarn, S. (2022). Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC medical research methodology*, 22(1), 281.
- [14] Chen, X., Duan, S., Li, S., Liu, D., & Fan, H. (2023). A method of network public opinion prediction based on the model of grey forecasting and hybrid fuzzy neural network. *Neural Computing and Applications*, 35(35), 24681-24700.
- [15] Zhang, M., Zheng, R., Chen, J., Zhu, J., Liu, R., Sun, S., & Wu, Q. (2019). Emotional component analysis and forecast public opinion on micro-blog posts based on maximum entropy model. *Cluster Computing*, 22, 6295-6304.
- [16] Dong, X., Lian, Y., Tang, X., & Liu, Y. (2020). The damped oscillator model (DOM) and its application in the prediction of emotion development of online public opinions. *Expert Systems with Applications*, 148, 113268.
- [17] Chen, X. G., Duan, S., & Wang, L. D. (2017). Research on trend prediction and evaluation of network public opinion. *Concurrency and Computation: Practice and Experience*, 29(24), e4212.
- [18] Su, Q., Yan, S., Wu, L., & Zeng, X. (2022). Online public opinion prediction based on a novel seasonal grey decomposition and ensemble model. *Expert Systems with Applications*, 210, 118341.
- [19] Uthirapathy, S. E., & Domic, S. (2020). Real-Time Opinion Prediction Method for Emergency Public Events in Social Media Networks Using Opinion Hit Matrix. *Rev. d'Intelligence Artif.*, 34(4), 507-514.
- [20] Yang, G. R., Wang, X., Ding, R. X., Cai, J. T., Xu, J. D., & Herrera-Viedma, E. (2024). A method of predicting and managing public opinion on social media: An agent-based simulation. *Information Sciences*, 674, 120722.
- [21] Xu, L., Qiu, J., & Zhai, J. (2023). Trend prediction model of online public opinion in emergencies based on fluctuation analysis. *Natural Hazards*, 116(3), 3301-3320.
- [22] Mu, G., Liao, Z., Li, J., Qin, N., & Yang, Z. (2023). IPSO-LSTM hybrid model for predicting online public opinion trends in emergencies. *PLoS One*, 18(10), e0292677.
- [23] Xie, J., Zhang, S., & Lin, L. (2022). Prediction of network public opinion based on bald eagle algorithm optimized radial basis function neural network. *International Journal of Intelligent Computing and Cybernetics*, 15(2), 260-276.
- [24] Khan, S., Raza, S. H., Ilyas, M., Shah, A. A., Zaman, U., Ogadimma, E. C., & Sattar, S. (2025). Hybrid model of machine and deep learning to analyze Twitter data and prediction of online public opinion: revisiting agenda-setting implications. *Information Discovery and Delivery*.
- [25] Tang, C. (2025). A Deep Learning-Based Study on Predicting Changes in Data-Driven Opinion Dynamics in Social Media. *J. COMBIN. MATH. COMBIN. COMPUT*, 127, 1387-1409.
- [26] Imen Jarraya, Safa Ben Atitallah, Fatimah Alahmed, Mohamed Abdelkader, Maha Driss, Fatma Abdelhadi & Anis Koubaa. (2025). SOH-KLSTM: A hybrid Kolmogorov-Arnold Network and LSTM model for enhanced Lithium-ion battery Health Monitoring. *Journal of Energy Storage*, 122, 116541-116541.
- [27] Xinyu Guan, Hanyu Chen, Yali Liu, Ziwei Zhang & Linhong Ji. (2025). Predicting ground reaction forces and center of pressures from kinematic data in crutch gait based on LSTM. *Medical Engineering and Physics*, 139, 104338-104338.
- [28] Zhe Xu, Changyin Zhao, Xin Ning, Mengyao Qin, Zhen Zhang & Fuquan Nie. (2025). Configuration optimization for a plate-fin heat exchanger combining Taguchi method and multi-objective grey wolf optimizer. *Case Studies in Thermal Engineering*, 69, 106045-106045.
- [29] Musab Alataiqeh, Hu Shi, Qiangqiang Qu, Xuesong Mei & Haitao Wang. (2025). Thermal error modeling of slant bed CNC lathe spindle based on BiLSTM with data augmentation and grey wolf optimizer algorithm. *Case Studies in Thermal Engineering*, 70, 106090-106090.
- [30] Somak Maitra & Dimitris N. Politis. (2024). Prepivot Augmented Dickey-Fuller Test with Bootstrap-Assisted Lag Length Selection. *Stats*, 7(4), 1226-1243.