A Method for Simulating the Propagation of Network Hot Topics Based on Combined SEIRS-ARIMA Model

Leyang Yu, Jingzu Xia, Yuli Wang

School of Mathematics and Statistics, Southwest University, Chongqing, China, 400715

Keywords: Internet Hot Topics, SEIRS Model, ARIMA Model.

Abstract: The spread of hot topics on the internet can easily attract public attention and discussion. However, the spread of rumours can cause social instability, while positive media guidance can generate economic value. Therefore, developing a model that can understand the propagation of popular topics on the internet is a worthwhile pursuit. In this study, a hybrid model is developed by combining the SEIRS (Susceptible-Exposed-Infected-Recovered-Susceptible) infectious disease model to simulate the overall trend and the ARIMA (Autoregressive Integrated Moving Average) model to simulate noise, to explain internet hot topic propagation. Using the wordle game as a case study, the model exhibits a notable capability in conforming to the general pattern and the majority of the minor variations, while demonstrating a high degree of interpretability.

1. Introduction

The global internet has emerged as a crucial platform for accessing information and engaging in interactions, while hot topics on the internet have attracted public attention and discussions. These topics are characterized by diverse communication methods, fast updates, and strong participation and interaction. However, the prevalence of malicious topics poses a severe threat to social stability and leads to significant losses [1]. In contrast, positive media guidance can generate economic value [2]. Therefore, developing a model that comprehends the propagation of popular topics on the internet is a worthwhile pursuit.

Traditionally, the proliferation of popular topics on the internet has been treated as a time series problem and modelled using ARIMA time series models [3]. However, recent developments in neural network research have established LSTM models as a promising alternative for prediction. Nevertheless, these models lack interpretability and fail to capture the internal patterns in depth [4].

Alternatively, the process of information diffusion is a continuous result of people influencing each other over some time [5], so infectious disease models can be employed to model the propagation of web topics. Just as people contract infectious diseases by coming into contact with infectious sources, web users can be seen as potentially infected individuals who transmit information by browsing, sharing, and commenting. Therefore, the infectious disease model provides an appropriate framework to address such issues. The SIR model proposed by Kermack forms the basis of the infectious disease model, which included three states: susceptible state, infection state, and recovery state [6]. Subsequently, the SIS model [7], which can be infected again, and the SEIR model [8], which incorporates the exposed state, have emerged to handle more complex situations[9].

In this paper, we introduce the infectious disease model to investigate the spread of hot topics on the internet. We propose the SEIRS model, which can be repeatedly infected by integrating the characteristics of internet transmission. We also employ the ARIMA model to explain the noise that cannot be accounted for by the SEIRS model and analyze the spread of the popular game Wordle in 2022.

2. The fundamental of SEIRS and ARIMA

2.1 The structure of the SEIRS model

The SEIRS (Susceptible-Exposed-Infected-Recovered-Susceptible) model, a well-established and robust traditional infectious disease model is a non-linear kinetic model that enables a comprehensive study of the transmission rate, the number of infections, and the temporal dynamics of infectious diseases and provides valuable guidance for the development of effective strategies for the control and prevention of infectious diseases [10].

The SEIRS model classifies individuals according to their infection status, and when introducing the SEIRS model to the propagation of internet hot topics, we can also level these statuses to the internet population. The susceptible group (denoted S) includes individuals who are not yet infected with the disease and remain healthy, and represents a large portion of the population, referring to the number of people who will be affected by the issue of the spread of the internet hot topic. The exposed group (denoted E) includes individuals who are in the incubation period of the disease and refers to individuals who are already aware of the hot topic but have not yet been spread. The infected group (denoted I) includes individuals who have been diagnosed with the disease, referring to individuals who are already affected by the hot topic and are contagious. The recovery group (denoted as R) includes those individuals who have been removed from the infected group and refers to those individuals who have dropped out of the hot topic.

The SEIRS model uses several parameters to describe the evolution between each group, and these parameters also have profound practical implications in the propagation of network hot topics. The SEIRS model adjusts some features of the fitted process by controlling these parameters to better simulate the real infection process.

The transmission coefficient, denoted by β, indicates the average number of infected individuals in the population over a given period. In the context of the spread of a network hot topic, this refers to the average number of people already affected by the hot topic who will spread. The higher the value of β, the greater the chance of transmission. the interrelationship between β, S, and I can be described by the following equation:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} + \xi R, \tag{1}$$

where t represents time, and N represents the population size, the sum of individuals in the susceptible, exposed, infectious, and recovered groups. In addition, the rate of immunity loss in recovered individuals is expressed by ξ, the details of which are discussed below. Furthermore, the rate of transition to the infection period in exposed individuals is denoted by σ. This rate is quantified as the inverse of the mean time of the incubation period. It reflects the rate at which the disease's incubation period ends and the individual becomes infectious. In the context of internet hot topic formation and transmission, it refers to the rate at which individuals who have heard about the hot topic will become affected. Individuals can leave E by entering the infected group, as determined by the following equation:

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E. \tag{2}$$

Once an individual is infected, there is a period before they enter the recovery group, and the rate at which this occurs is denoted as γ. In other words, γ indicates the number of individuals recovering from infection per unit of time, which is the rate at which those affected by the hot topic are withdrawn from it. The following equation depicts the correlation between γ, I, and E:

$$\frac{dI}{dt} = \sigma E - \gamma I. \tag{3}$$

After recovery, individuals may also shift from the recovery group to the susceptible group, which is reflected as re-following during the formation and dissemination of hot topics on the internet, a ratio ξ to describe that it is usually relatively small. We have this equation:

$$\frac{dR}{dt} = \gamma I - \xi R. \tag{4}$$

By integrating the equations presented above, we derive a system of differential equations. Upon solving the system, we can obtain simulated values that correspond to the modelled phenomena. Figure 1 illustrates the basic process of the SEIRS model.

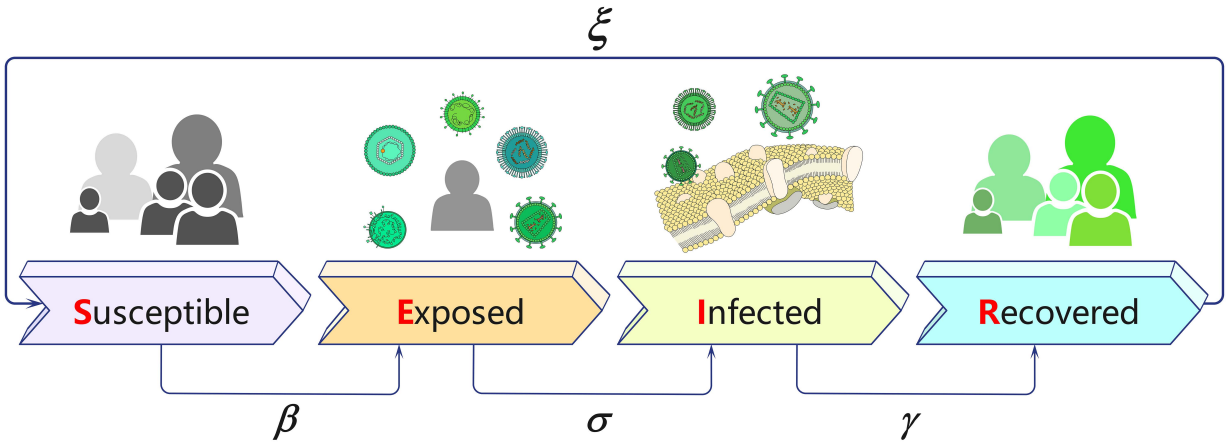

Figure 1 The basic process of the SEIRS model

2.2 The structure of the ARIMA model

Time series forecasting frequently utilizes the ARIMA (Autoregressive Integrated Moving Average) model as an approach that models the data series of the forecast object as a stochastic process and utilizes a specific mathematical model to approximate the underlying series. The ARIMA model has three basic components, each of which has a parameter (p,d,q). The first part is the "Autoregressive" (AR) part, which describes the relationship between the current values and the historical data, determined by p, and represents the number of lags of the time series data itself used in the model. The formula is as follows:

$$\left(1 - \phi_1 \mathcal{B} - \cdots - \phi_p \mathcal{B}^p\right) y_t = c + \varepsilon_t \tag{5}$$

where $\mathcal{B}$ is the lag operator, y_t is the time series data, c is a constant and ε_t is the white noise.

The second part is the "integrated" part, which is the number of differences needed to smooth the time series, denoted by d. Usually, one time is sufficient. After adding the second part, the above equation takes the following form:

$$\left(1 - \phi_1 \mathcal{B} - \cdots - \phi_p \mathcal{B}^p\right)(1 - \mathcal{B})^d y_t = c + \varepsilon_t \tag{6}$$

The third component is the "moving average" component, which focuses on the accumulation of error terms, and it is usually controlled by the parameter q, which represents the number of lags of the error terms used in the model. This relationship can be expressed by the following equation:

$$y_t = c + \left(1 + \theta_1 \mathcal{B} + \theta_q \mathcal{B}^q\right)\varepsilon_t \tag{7}$$

Combining these three components is the ARIMA model, which has the following equation:

$$\left(1 - \phi_1 \mathcal{B} - \cdots - \phi_p \mathcal{B}^p\right)(1 - \mathcal{B})^d y_t = c + \left(1 + \theta_1 \mathcal{B} + \theta_q \mathcal{B}^q\right)\varepsilon_t \tag{8}$$

2.3 Combined SEIRS-ARIMA model

In this paper, we introduce an innovative Combined SEIRS-ARIMA model that combines the SEIRS model with the ARIMA model.

While the SEIRS model is an ideal infectious disease model that can simulate and explain the similarity between the spread of hot topics on the internet and infectious diseases, it does not take into account other factors that affect the spread of information, such as external information, media coverage, and the psychological state of users. Therefore, we divide the spread of real hot topics into two parts: one part can be simulated and explained by the SEIRS model, which accounts for the majority of the spread, and the other part, which we refer to as "noise", is simulated using the ARIMA model. Our total model is the sum of these two parts.

3. Data processing and simulation results

3.1 Data preprocessing

In this subsection, we conduct a simulation of the dissemination of the popular game "Wordle" in 2022, utilizing data procured from the MCM contest. Use the variable "Number of reported results" in the dataset, which is the total number of scores that were recorded on Twitter that day, and the time we use starts with 2022-1-7, and ends with 2022-11-15. Given that the data was obtained via a web crawler from Twitter by the contest organizers, the possibility of unreliability could not be discounted. To ensure data veracity, we perform data cleaning measures to counteract any potential outliers and absent values. After a meticulous evaluation of the dataset, we find no instances of absent values. However, our exploratory data analysis detects an atypical event with substantial fluctuations on November 30, 2022, which we deem highly improbable. To mitigate this anomaly, we approximate the values at this time through the utilization of the average values of the preceding and following days. This guarantees that our simulations are executed on a reliable and uniform dataset. After completing the data cleaning, the timing diagram is shown in Figure 2.

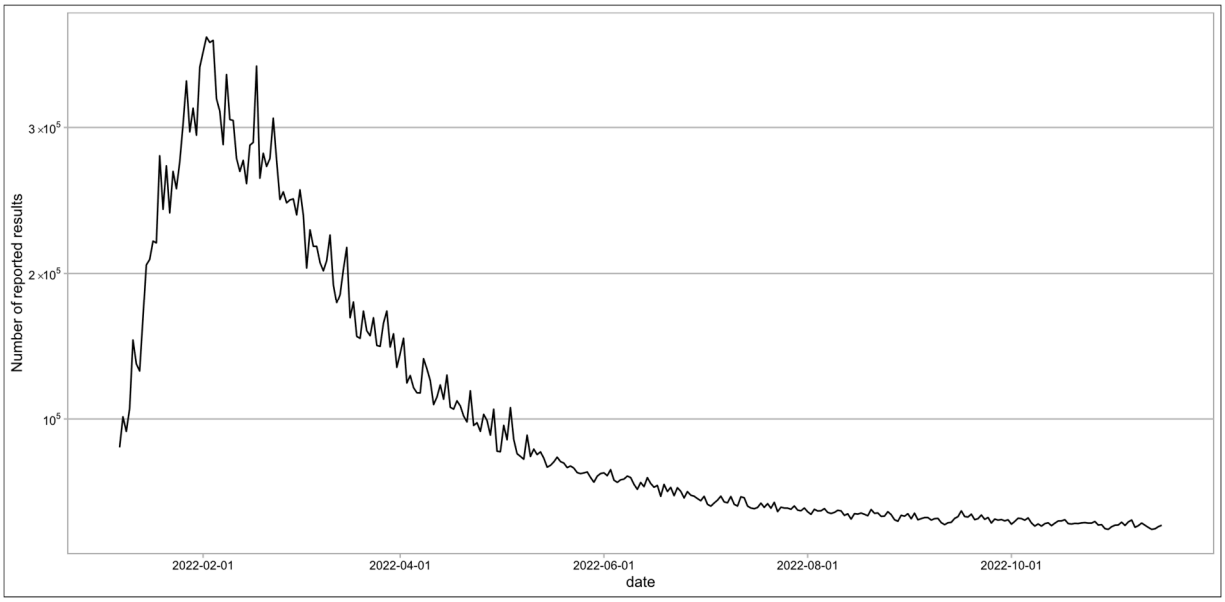

Figure 2 Time-series diagram of the studied variable

3.2 Simulation results

As a matter of fact, the spread of the network will go through different periods, and in different periods, they have different performance characteristics. This can be seen in Figure 3, which is plotted by the data after first-order differencing and can reflect the dramatic degree of data change.

From Figure 3, we observe that the spread of internet hot topics exhibits different degrees of fluctuation at different stages, as delineated by the red dashed line. Specifically, we observe more pronounced changes in the first half of the delineated line, with significantly less fluctuation in the second half. We attribute this phenomenon to the gradual decay of popularity levels after the initial explosion period. Accordingly, we propose a segmentation approach to the study of internet hot topic propagation, which involves dividing the spreading process into two distinct phases: the explosion

period and the gradual levelling-off period. Given that these two phases exhibit divergent propagation characteristics, we contend that a segmentation approach is a reasonable and effective means of tackling the problem of internet hot topic propagation.

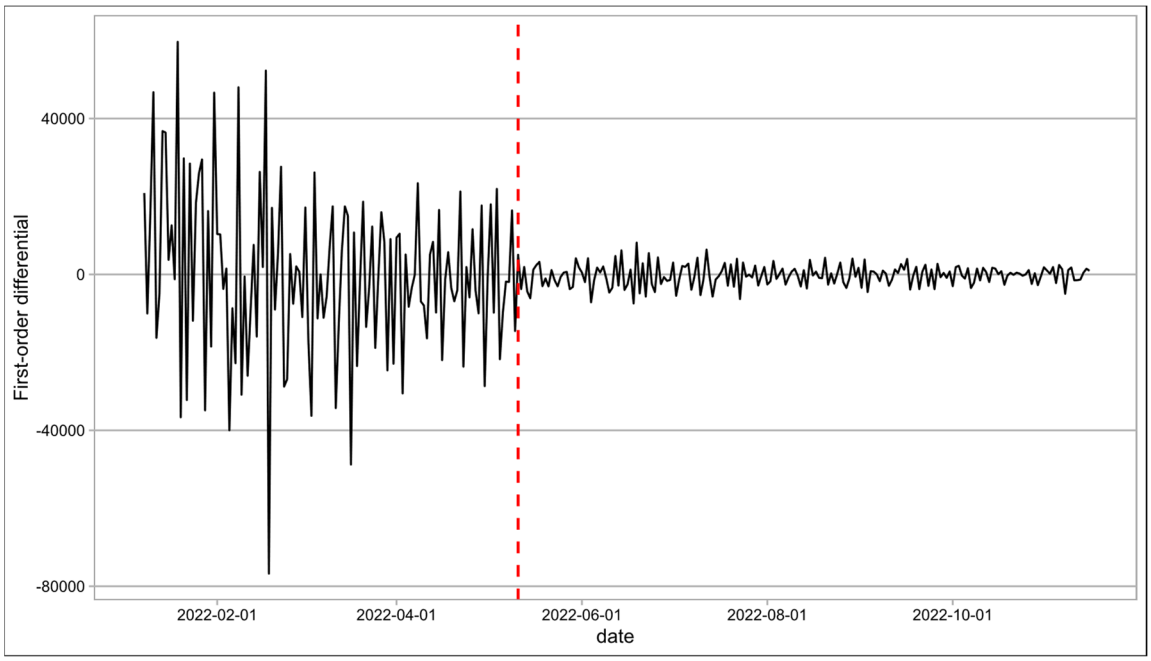

Figure 3 First-order difference plots of the studied variable

In the context of internet hot topic outbreaks, the SEIRS model is expected to exhibit a higher infection rate, owing to the contagious nature of such issues. Furthermore, given that many individuals may engage with the hot topic out of novelty or transient curiosity, the model should also reflect a higher recovery rate. Additionally, the rate of reversion to the hot topic is likely to be elevated at this stage, due to social factors such as herd mentality. In the simulation of wordle game popularity propagation, the cut-off point is set to May 11, 2022, with the outbreak period in the previous period and the smooth period in the latter period. The corresponding SEIRS model parameters are set as $\beta = 0.3$, $\sigma = 0.15$, $\xi = 0.0015$, and $\gamma = 0.019$ for the outbreak period, and $\beta = 0.2$, $\sigma = 0.15$, $\xi = 0.001$, and $\gamma = 0.015$ for the smooth period to obtain the simulated images shown in Figure 4.

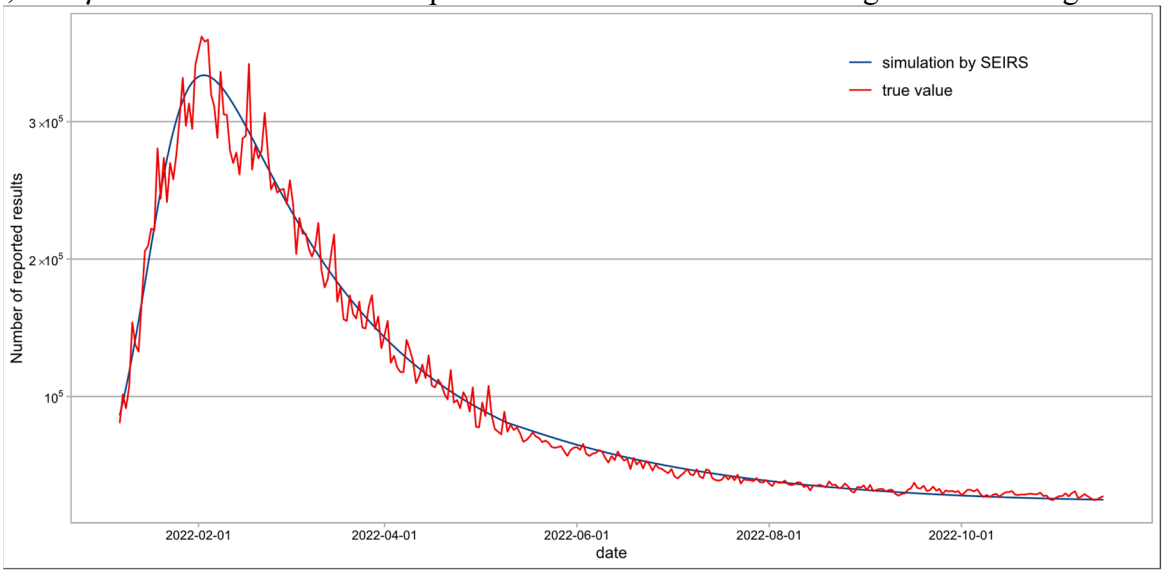

Figure 4 Simulation curve of the SEIRS model

Figure 4 shows that the SEIRS model can effectively model and explain the similarity between the spread of hot topics on the internet and infectious diseases. However, it is crucial to acknowledge that other factors can influence information propagation, which the SEIRS model does not account for.

To address this limitation, we utilize an ARIMA model to fit the noise. And there is a fact that the propagation pattern varies across different stages of internet hot topic spread. To effectively capture these variations, we develop segmentation parameters for the ARIMA model as ARIMA (0,0,2), and ARIMA (2, 1, 0). Using these two models, we obtain simulations of the noise, as shown in Figure 5.

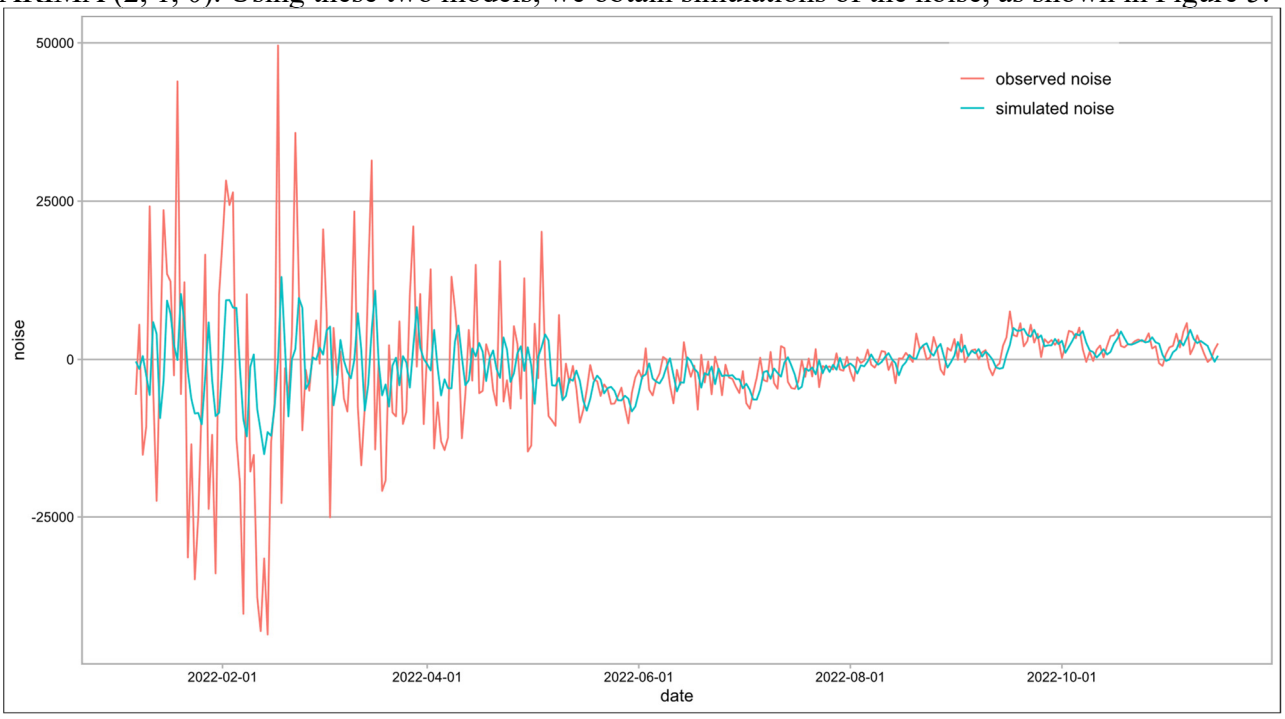

Figure 5 Noise simulation obtained from the ARIMA model

Figure 5 reveals that the simulation is successful in capturing the majority of the noise trends, particularly during the gradual stabilization phase. However, significant disparities are observed in the popularity burst period. This discrepancy may arise due to the data collection method, as the data is obtained through web crawlers, which may lack reliability. Using more reliable data sources, such as official popularity change data, may result in improved simulation outcomes for our model.

Figure 6 Simulation curves of the SEIRS-ARIMA model

By associating the SEIRS model with the ARIMA model, we obtain our simulation of the dissemination process of hot topics on the internet shown in Figure 6. From Figure 6, it can be observed that the simulation has been successful in capturing most of the real characteristics, as well as the small

fluctuations. However, there are some points in the outbreak period where there are relatively large discrepancies between the simulated and actual values. as described in the simulation of the noise. This discrepancy may be attributed to the limitations of the data source, which was obtained through web crawling, resulting in some incomprehensive data. It is expected that the model would perform better with more reliable and comprehensive data, such as official popularity change data.

4. Conclusions and outlooks

Through this study, we successfully simulate the propagation of online hot topics using the joint SEIRS-ARIMA model and achieve good simulation results. Our model has good explanatory power and can clearly describe the propagation pattern and influencing factors of online hot topics. Our study is based on the infectious disease transmission model SEIRS and applies it to model the transmission of hot topics on the internet. This modelling approach is consistent with the basic perception of "human-to-human" transmission of internet hot topics and effectively portrays the transmission process of internet hot topics by introducing factors such as exposure, infection, recovery, and susceptible people. In addition, we also use the ARIMA model to fit other noises in the transmission process, which further improves the accuracy of our model. Taken together, our joint SEIRS-ARIMA model not only simulates the spread of online hot topics well but also considers various factors in the spread process, which provides a powerful tool and support for further research on the spread of online hot topics.

Although we apply the SEIRS-ARIMA model in this study, it is reasonable to suggest that the concept of combining different models can be extended to a broader range of applications since other models can be integrated with the SEIRS model easily. For example, the ARIMA model can be replaced by other models to predict the noise generated by SEIRS in different contexts. In this study, we do not use LSTM due to concerns about overfitting for small sample sizes, but for large samples, the proposed approach can be extended to replace ARIMA with LSTM [11]. Additionally, while the SEIRS model is used in this study, other models can be incorporated depending on the specific problem, For example, the SAIS model [12]. Examining different combinations of models for various problems may lead to better outcomes. Furthermore, the hot topic propagation on the internet is only divided into two stages in this study. In practical applications, there can be more than two stages, and dividing the propagation into different stages and setting the parameters specific to each segment may improve the simulation results of the model.

Finally, we present two application scenarios to demonstrate the practical use of the model. The spread of hot topics on the internet encompasses two main types of issues that attract greater attention. Firstly, the spread of rumours can lead to social disturbances, public safety concerns, and harm to public interests. Employing the theory of infectious diseases, it can be effectively controlled [13]. With the aid of the network hot topic dissemination model, a range of strategies can be implemented to modify corresponding model parameters and hence control the further spread of rumours. Secondly, there is the exploitation of network hot topics for profit-making purposes, such as in the spread of game popularity, where individuals are more concerned about rapidly increasing and maintaining long-term popularity. By utilizing the network hot topic dissemination model, a suite of techniques can be applied to enhance the spreading rate while reducing the recovery rate, thereby achieving the goal of long-term profitability.

References

[1] Gosling S D, Mason W. Internet research in psychology[J]. Annual review of psychology, 2015, 66: 877-902.

[2] Gruner R L, Vomberg A, Homburg C, et al. Supporting new product launches with social media communication and online advertising: sales volume and profit implications[J]. Journal of Product Innovation Management, 2019, 36(2): 172-195.

[3] Cai M, Luo H, Cui Y. A Study on the Topic-Sentiment Evolution and Diffusion in Time Series of Public Opinion Derived from Emergencies[J]. Complexity, 2021, 2021: 1-23.

[4] Liu P. Time Series Forecasting Based on ARIMA and LSTM[C]//2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022). Atlantis Press, 2022: 1203-1208.

[5] Kleinberg J. The convergence of social and technological networks[J]. Communications of the ACM, 2008, 51(11): 66-72.

[6] Kermack W O, McKendrick A G. A contribution to the mathematical theory of epidemics[J]. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 1927, 115(772): 700-721.

[7] Zhu L, Huang X. SIS model of rumor spreading in social network with time delay and nonlinear functions[J]. Communications in Theoretical Physics, 2019, 72(1): 015002.

[8] Heng K, Althaus C L. The approximately universal shapes of epidemic curves in the Susceptible-Exposed-Infectious-Recovered (SEIR) model[J]. Scientific Reports, 2020, 10(1): 19365.

[9] Brauer F, Castillo-Chavez C, Castillo-Chavez C. Mathematical models in population biology and epidemiology[M]. New York: springer, 2012.

[10] Mummert A, Otunuga O M. Parameter identification for a stochastic SEIRS epidemic model: case study influenza[J]. Journal of mathematical biology, 2019, 79: 705-729.

[11] Siami-Namini S, Tavakoli N, Namin A S. A comparison of ARIMA and LSTM in forecasting time series[C]//2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018: 1394-1401.

[12] Sahneh F D, Scoglio C. Epidemic spread in human networks[C]//2011 50th IEEE Conference on Decision and Control and European Control Conference. IEEE, 2011: 3008-3013.

[13] Wang J, Jiang H, Ma T, et al. Global dynamics of the multi-lingual SIR rumor spreading model with cross-transmitted mechanism[J]. Chaos, Solitons & Fractals, 2019, 126: 148-157.