IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites **SUBSCRIBE** **SUBSCRIBE** Cart Create Account Perso Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Journals & Magazines > IEEE Transactions on Cybernet... > Volume: 50 Issue: 7

Predicting COVID-19 in China Using Hybrid AI Model

Publisher: IEEE Cite This PDF

Nanning Zheng ; Shaoyi Du ; Jianji Wang ; He Zhang ; Wenting Cui ; Zijian Kang ; Tao Yang ; Bin Lou ; Yuting Chi ; Hong Lo… **All Authors** •••

156 Paper Citations **12000** Full Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Free

Abstract

Document Sections

I. Introduction

II. Framework of the Hybrid AI Model

III. Analysis of the Laws of Epidemic Transmission

IV. Prediction of the Development Trend of the Epidemic

V. Experimental Results

Show Full Outline

Authors

Figures

References

Citations

Keywords

Metrics

Downl PDF

Abstract: The coronavirus disease 2019 (COVID-19) breaking out in late December 2019 is gradually being controlled in China, but it is still spreading rapidly in many other countri... **View more**

Metadata

Abstract:

The coronavirus disease 2019 (COVID-19) breaking out in late December 2019 is gradually being controlled in China, but it is still spreading rapidly in many other countries and regions worldwide. It is urgent to conduct prediction research on the development and spread of the epidemic. In this article, a hybrid artificial-intelligence (AI) model is proposed for COVID-19 prediction. First, as traditional epidemic models treat all individuals with coronavirus as having the same infection rate, an improved susceptible–infected (ISI) model is proposed to estimate the variety of the infection rates for analyzing the transmission laws and development trend. Second, considering the effects of prevention and control measures and the increase of the public's prevention awareness, the natural language processing (NLP) module and the long short-term memory (LSTM) network are embedded into the ISI model to build the hybrid AI model for COVID-19 prediction. The experimental results on the epidemic data of several typical provinces and cities in China show that individuals with coronavirus have a higher infection rate within the third to eighth days after they were infected, which is more in line with the actual transmission laws of the epidemic. Moreover, compared with the traditional epidemic models, the proposed hybrid AI model can significantly reduce the errors of the prediction results and obtain the mean absolute percentage errors (MAPEs) with 0.52%, 0.38%, 0.05%, and 0.86% for the next six days in Wuhan, Beijing, Shanghai, and countrywide, respectively.

Published in: IEEE Transactions on Cybernetics (Volume: 50 , Issue: 7, July 2020)

More Like This

Page(s): 2891 - 2904

Date of Publication: 08 May 2020 ?

▸ **ISSN Information:**

PubMed ID: 32396126

▸ **Funding Agency:**

INSPEC Accession Number: 19687832

DOI: 10.1109/TCYB.2020.2990162

Publisher: IEEE

☰ Contents

SECTION I.
Introduction

The outbreak of the coronavirus disease 2019 (COVID-19), which quickly spread across the country, coincided with the spring festival period in China. In its primary stage of transmission, the COVID-19 was not effectively suppressed because of the extreme irregularity of the primary stage of the epidemic, the limited understanding of the new coronavirus by the medical community, and the lack of medical resources [1]. The COVID-19 can be transmitted from person to person, as officially confirmed on January 20, 2020 [2]. Therefore, all provinces and cities in China have implemented strong prevention and control measures, including the closure of the airport and railway stations in Wuhan on January 23, 2020, which are considered the strictest epidemic control measures in history. Public awareness of epidemic prevention has gradually increased because of these effective prevention and control measures. Presently, the number of new infections has decreased significantly. From February 3, 2020 to February 19, 2020, the number of new daily confirmed cases outside Hubei has dropped for 16 consecutive days; the number of new infections in Hubei has also been gradually decreasing since February 12, 2020, and the number of cured patients has increased. The epidemic prevention and control have achieved initial success in China, but in other countries and regions, especially in Europe, Iran, South Korea, the US, and Japan, the epidemic situation is still severe. Every country or region needs to develop targeted prevention and control strategies to control the epidemic effectively. Therefore, conducting research on the development and spread of epidemics is necessary. In the current case, analyzing the development law and predicting the trend of COVID-19 are crucial for effective prevention and control of this epidemic.

When a large-scale epidemic infectious disease emerges and a major public health emergency is initiated, people utilize epidemic models to analyze and predict the development trend of the disease and use the analysis results to guide the development of the prevention and control measures. The most widely used traditional epidemic models are susceptible–infected (SI), SI recovered (SIR), and susceptible–exposed–infected–recovered (SEIR) models [3]– [5], where “S,” “E,” “I,” and “R” denote the number of susceptible people, the number of people in the incubation period, the number of infectious cases, and the number of people who have recovered, respectively. SI, SIR, and SEIR models represent the relationship between I and S in the form of differential equations. These models have been successfully applied to the prediction of various diseases, such as Ebola and SARS, because of their strong disease prediction capabilities [6]– [10]. Given the severe situation of COVID-19, the analysis of changes in the number of new daily confirmed cases is particularly important for inferring the trend of an epidemic. Therefore, we need to focus on the impact of the trend of new infections on the spread of an epidemic. Furthermore, the influence of cure and mortality rates on the trend of the epidemic are not considered in this article because both parameters have no direct relationship with the number of new daily confirmed cases.

Traditional epidemic models analyze the infection rate based on the dynamic change in the number of infections and subsequently predict the spread and development trend of the epidemic. However, these models consider that all individuals with coronavirus have the same infection rate. Their prediction results can only provide general trends and, thus, have limitations. The government's prevention and control measures have a significant impact on the containment of the development trend of the epidemic, and transparent reporting of the epidemic, implementation of prevention and control measures, and reinforcement of residents' prevention awareness have accelerated the containment of the virus. Evidently, epidemic data alone are insufficient to achieve accurate prediction. We must build a data-driven epidemic model for public health emergencies. By using news information features, we can overcome the limitation of traditional epidemic models that use only a single factor, further improve the accuracy of model prediction, and verify the effectiveness of the government's prevention and control strategies.

To deal with this problem, the long short-term memory (LSTM) network with the natural language processing (NLP) module is introduced into our epidemic model to update the infection rate and further improve the predictive accuracy of the model. LSTM is a classic recurrent neural network (RNN) proposed by Hochreiter and Schmidhuber [11]. LSTM can effectively alleviate gradient explosion and gradient disappearance during the training procedure by introducing the constant error carousel unit. Compared with traditional RNN [12], LSTM exhibits better performance in capturing the long-term dependencies of sequences and is therefore suitable for the classification, processing, and prediction of long-sequence data [13]– [16]. In recent years, LSTMs have been widely used in various tasks, such as NLP [17]– [20]; image generation [21], [22]; and video analysis [23], [24].

This article focuses on the analysis of the infection rate of individuals with coronavirus, models the ability of viruses to infect susceptible people according to different periods after infection, and proposes an improved susceptible–infected (ISI) model. Based on the proposed ISI model, the hybrid artificial-intelligence (AI) model embedded the NLP module and LSTM network for predicting the COVID-19 in this article, and it introduces the important information of the great efforts led by the central government and local governments as well as the massive support participation from the public into the prediction calculation process. Furthermore, this article analyzes the development of the epidemic based on the proposed hybrid prediction model and predicts the trend of the epidemic. The experimental results obtained based on the epidemic data of several typical provinces and cities in China show that the proposed hybrid model can provide a basis for estimating the law of virus spread, and achieve more accurate and robust performance compared with the traditional epidemic models. Moreover, the prediction results of our hybrid AI model with the introduction of news information are more in line with the actual epidemic development trend, which demonstrates that the openness, transparency, and efficiency of data releasing are very important for establishing a modern epidemic prevention system.

The remainder of this article is organized as follows. Section II introduces the framework of the proposed AI model. Section III proposes the ISI epidemic model to analyze the laws of epidemic transmission. Section IV gives the NLP-based LSTM model for precise prediction. Section V provides the experimental results based on the epidemic data of several typical provinces and cities in China. The conclusion is provided in Section VI.

SECTION II.
Framework of the Hybrid AI Model

In existing epidemic models, the infection source of new daily confirmed patients in the future consists of those with coronavirus that are not quarantined. Therefore, most epidemic models regard the number of patients who are infected but not quarantined as the base, and then multiply the estimated infection rate to predict the number of new daily confirmed cases [25]– [27]. However, the infection rate of individuals with coronavirus varies at different time intervals of infection [28]. Traditional epidemic models treat all individuals with coronavirus as having the same infection rate and are therefore unable to reflect the evolution trend of an epidemic. Under prevention and control measures, most new confirmed cases at this moment are infected by the [illegible] establishment of the epidemic model in this article because these cases have no direct impact on the number of new confirmed cases. Based on this assumption, we propose an ISI epidemic model that uses a retrospective approach and a grouped multiparameter method. The basic principle of the retrospective approach is to use the ratio of the number of new confirmed cases at time t to the cumulative number of new confirmed cases over different time scales before time t to calculate the infection rate and establish an epidemic model. Furthermore, the importance of different time scales to the new confirmed cases at time t is analyzed in accordance with the prediction result of the model. Grouped multiparameter factors, which determine the impact of confirmed cases at different times before time t on the confirmed cases at time t , are used to the ISI model to quantify the infection rate of infected cases at different periods. Then, the improved model is used for analyzing the development law of infectious diseases.

In addition, the LSTM network is used to estimate the infection rate deviation of the epidemic model and is combined with the proposed ISI model to estimate the number of infected cases. To consider the influence of government control measures, the media's transparent reports, and the increase in public awareness regarding epidemic prevention, this article uses pretrained NLP models to extract features from relevant news of various provinces and cities. The extracted features are subsequently combined with the LSTM network to correct the deviation of the infection rate estimated by the ISI model, which could predict the number of infected cases based on the transmission laws and development trend. The proposed framework is shown in Fig. 1.

Fig. 1.
Hybrid AI model for COVID-19 prediction by using all historical data.

SECTION III.
Analysis of the Laws of Epidemic Transmission

Traditional epidemic models deem that the number of new infectious cases is related to the number of people who are infected and susceptible, but these models still lack an in-depth analysis. People undergo different infection cycles for different infectious diseases [29]. The time distribution of the infectious sources of new daily confirmed cases must be determined to investigate the infection law of an epidemic. The purpose of this article is to analyze the spread laws and development trend of an epidemic by modeling new confirmed data. However, cure and mortality rates are not directly related to the number of new confirmed cases, so they are not considered in this article.

The observation period for COVID-19 is 14 days [30], so we can posit that almost all new daily confirmed cases are infected by patients confirmed in the past 14 days. Most of the patients under investigation have been quarantined, observed, and tested with a nucleic acid reagent. Patients need to obtain at least two positive results before being confirmed as positive for COVID-19, so we can infer that most of the confirmed patients have been quarantined at least three days prior to the confirmation and are unable to infect others, which means that most of the confirmed patients cannot be infected by another confirmed case who was confirmed 11 days ago. Therefore, for each day t , this article examines the infection rate of new daily confirmed cases in the past ten days relative to the confirmed cases of day t . For an enhanced analysis, the following symbols are defined: $S(t)$ represents the number of susceptible persons on day t , $I(t)$ represents the cumulative number of confirmed cases of day t (inclusive), and $\Delta I(t) = I(t) - I(t-1)$ represents the number of new confirmed cases on day t .

To obtain a comprehensive understanding of the impact of the infected cases on subsequent infected persons, we need to determine the time interval in which the confirmed cases are most likely to infect the new daily confirmed cases at day t . Then, according to the difference between them, we can determine the laws of transmission. Therefore, in this article, we consider that the patients confirmed on day t are infected by a confirmed person from day $t-1$ to day $t-10$. To determine at which stage the current new daily confirmed cases are infected at a high infection rate, we use the retrospective method to analyze the time laws of epidemic transmission in the past few days. We develop an improved multiparameter epidemic model for the past ten days by applying it to several infection periods and conduct an in-depth analysis of the time laws of the transmission of COVID-19. The framework of the ISI epidemic computational model is shown in Fig. 2.

Fig. 2.
ISI model.

A. Correlation Model of Infection Rate by Using the Retrospective Method

Traditional epidemic models generally deem that the confirmed cases on a certain day originate from confirmed cases in the past few days. Most of the epidemic models in previous research [31], [32] are based on a fixed number of days and assume that the transmission of an epidemic is affected by previous k days. However, these models lack an in-depth analysis of how epidemics are transmitted. According to the general laws of the development of epidemics, compared with the patients confirmed at the adjacent time (e.g., day $t-1$) on day t , early confirmed patients (e.g., day $t-5$) are more likely to affect patients confirmed on day t . Therefore, by modeling on the infection rate of the cumulative number of confirmed cases in the past k days relative to the confirmed cases on day t , the laws of COVID-19 transmission can be obtained with enhanced macroscopic guiding significance for the overall trend estimation of epidemic development.

We use the retrospective method to analyze the influence of cumulative confirmed cases at different times on the estimation of the infection rate. We examine the infection rate of patients in different regions at different time intervals and investigate whether the current new confirmed cases are infected by the cumulative number of confirmed cases in the past k days. The equation is given as follows:

$$I(t) = I(t-1) + \beta_1(t,k)\sum_{i=1}^{k}\Delta I(t-i), \quad k = 1,2,\ldots,10 \tag{1}$$

View Source

where $\beta_1(t,k) = \Delta I(t)/\sum_{i=1}^{k}\Delta I(t-i)$ is the infection rate of the cumulative number of confirmed cases from day $t-k$ to day $t-1$ relative to the new confirmed cases on day t . It reflects the relationship between the number of new confirmed cases $I(t)$ on day t and the number of new confirmed cases $\sum_{j=1}^{k}\Delta I(t-k)$ in the past k days.

First, the purpose of determining $\beta_1(t,k)$ is to find out the relatively stable relationship between $\Delta I(t)$ and $\sum_{i=1}^{k}\Delta I(t-i)$, that is, to analyze the impact of the cumulative number of confirmed cases in the past k days on the number of new confirmed cases on day t . The aim is to make the laws of epidemic spread reasonably interpretable and provide support for follow-up work. Second, we can obtain the parameter $\hat{\beta}_1(t,k)$ of each province and the entire country according to (1). Given that the infection rate of epidemics changes exponentially, this article uses the exponential function $L(t) = a \times e^{-bt}$ to fit $\hat{\beta}_1(t,k)$, which can estimate the epidemic spread. In the formula, a and b are the parameters of the exponential function, and $a, b > 0$. Finally, patients have a strong infection rate because they cannot be effectively quarantined during the incubation period. Therefore, this section estimates $\beta_1(t,k)$ by gradually increasing the value of k , and the number of new confirmed cases at an earlier time point is gradually introduced into the model. Then, in accordance with the predictions of the model, we can analyze whether the new confirmed cases at each time will infect the new confirmed patients on day t . The evolution laws of patients in different time intervals in the process of epidemic transmission can also be obtained.

B. Influence Model of Infection Rate With Grouped Multiparameters

This article considers that infected cases cannot infect the number of susceptible people after being quarantined due to the strict control and quarantine measures. Therefore, the new confirmed cases on day t are most likely to be infected by the new confirmed cases in the past k days. A close relationship exists between the infection rate and the time of infection of patients [33]. Therefore, the new confirmed cases may have different infection rates for the new confirmed cases on day t at different times in the past k days. From (1), we estimate that the most possible infection time dates back to the recent several days. We further analyze this difference and quantify the contribution of new confirmed cases at different times to the infection rate at time t by giving different weights to the number of new confirmed cases each day from day $t-k$ to day $t-1$. Then, we estimate the infection rate by using the weighted cumulative confirmed number, which is adopted to model the epidemic.

To simplify the model, adjacent two days are regarded as a propagation unit, and the same weight α_i is assigned. Multiparameter epidemic modeling is then carried out, as shown in (2). The model avoids the drastic change in weight caused by single data abnormality, thus making the model more robust; at the same time, it reduces the search space of the weight and the complexity of the model

$$\begin{aligned} I(t) =& I(t-1) \\ &+ \beta_2(t,k)\sum_{i=1}^{k/2}\left(\alpha_i\left(\Delta I(t-2i+1) + \Delta I(t-2i)\right)\right) \end{aligned} \tag{2}$$

View Source

where $2\sum_{i=1}^{k/2}\alpha_i = 1$

Based on the above epidemic model, the model proposed in this section comprehensively considers the difference in the infection rate of new confirmed cases in the past k days relative to the new confirmed cases on day t to study the transmission correlation between the cumulative number of confirmed cases in the past k days and the number of the new cases on day t . First, several groups of different weights α_i are initialized randomly, and a multiparameter epidemic model is established by (2). The better the prediction result of the model, the better the corresponding weights reflect the real infection law. Finally, the infection rate with a great contribution to the virus infection can be inferred by comparing the weights assigned to different time points.

We obtain the relationship between the new confirmed cases on day t and the new confirmed cases on days $t-10$ to $t-1$ through the value of α_i on the basis of (1) and (2). However, too few parameters can cause underfitting [i.e., (1)], while too many parameters can easily cause overfitting [i.e., (2)]. Therefore, we further balance the number of parameters based on the above results. The set of days $\{t-i|i=1,2,\ldots,10\}$ is divided into two groups, where the set of days with a greater impact on the new confirmed cases on day t is recorded as set A , and the remaining days are recorded as set B . Set A is given weight γ_1 , and set B is given weight γ_2 , as in

$$I(t)=I(t-1)+\beta_3(t)\left(\gamma_1\sum_{t_1\in A}\Delta I(t_1)+\gamma_2\sum_{t_2\in B}\Delta I(t_2)\right) \tag{3}$$

View Source

where $\gamma_1|A|+\gamma_2|B|=1$, $|\cdot|$ denotes the number of elements in a set. We calculate the infection rate according to (3).

C. Data Preprocessing Method Based on the Proposed Model

The diagnostic criteria for patients at the beginning of the outbreak of COVID-19 changed throughout the country due to insufficient medical resources and limited understanding of the clinical signs of the novel coronavirus. These factors led to the presence of considerable noise in the epidemic data of all provinces. Hubei Province incorporated clinical diagnosis into the diagnostic criteria after the fifth edition of the treatment and diagnosis plan was released on February 12, 2020. This clinical diagnosis caused the new daily confirmed cases of Wuhan to surge to 13436 on that day. These abnormal and noisy data points bring great difficulties to subsequent modeling.

Data cleaning (i.e., removing abnormal data points) and the interpolation-based method are two widely applied approaches to deal with anomalous data points. However, these methods have many drawbacks. Data cleaning causes serious data loss and reduces the accuracy of the overall trend estimation of the epidemic model because the time scale of the epidemic data is extremely small. Meanwhile, although the interpolation-based method does not cause data loss, it loses the dynamic evolution laws of abnormal dates and affects the accuracy of short-term parameter estimation. Therefore, most of the new daily confirmed cases from abnormal data points are missed diagnoses from the early stage of the epidemic. Ignoring this number of patients will force the model to be too optimistic about the epidemic status of the early stage of the outbreak and will affect the modeling of subsequent evolution laws. For the abnormal data points near February 12, 2020, this article proposes a “data balance” method based on the epidemic model as a data preprocessing module to reduce the impact of changes in diagnostic criteria.

First, the data before February 12, 2020 are applied to build an epidemic model to predict the number of new daily confirmed cases on the anomaly dates. Second, the difference between all actual data points and the prediction results is summed up as the number of early missed patients. Third, these patients are evenly divided into abnormal and normal dates to achieve “trend balance” of the overall data. The implementation details are as follows.

1) Let the date with abnormal data start at t_s and end at t_e . Use $I(t_0)\cdots I(t_s)$ to establish an ISI epidemic model and predict the number of new daily confirmed cases on abnormal dates $\Delta\hat{I}(t_s)\cdots\Delta\hat{I}(t_e)$

2) Calculate the total number of missed diagnoses M and the cumulative number of new daily confirmed cases of the early stage N

$$\begin{aligned} M &= \sum_{t=t_s}^{t_e} \left(\Delta I\left(t\right) - \Delta \hat{I}\ \left(t_s\right)\right) \\ N &= \sum_{t=t_0}^{t_s-1} \Delta I\left(t\right) + \sum_{t=t_s}^{t_e} \Delta \hat{I}\ \left(t\right). \end{aligned} \tag{4}$$

View Source

3) Let $\alpha = M/N$. Then, the rebalanced data before t_e can be obtained by the following equation:

$$\begin{cases} \Delta I'\left(t\right) = \left(1+\alpha\right) \Delta I\left(t\right), & t = t_0, \ldots, t_s - 1 \\ \Delta I'\left(t\right) = \left(1+\alpha\right) \Delta \hat{I}\ \left(t\right), & t = t_s, \ldots, t_e. \end{cases} \tag{5}$$

View Source

The data balance preprocessing method has two main advantages.

1) The evolution trend of $\beta(t)$ will not be affected according to the calculation method [i.e., (1) – (3)] of the infection rate $\beta(t)$ if the numbers of the new daily confirmed cases before t_s are increased α times.

2) The number of new daily confirmed cases $I(t)$ before and after the anomaly date can maintain its evolution trend after all the data points before t_e have been enlarged; therefore, the long-term fitting result of $\beta(t)$ becomes increasingly stable. We select t_s as February 12, 2020 and t_e as February 13, 2020.

SECTION IV.

Prediction of the Development Trend of the Epidemic

The epidemic model can predict the spread of infectious diseases well but does not consider other factors, such as prevention and control measures, which prevent the spread of infectious diseases. Therefore, new mechanisms need to be introduced to update the parameters in the epidemic model. The LSTM network can be used to model hidden variables (e.g., number of potentially infected people) and is often utilized for data prediction. However, experiments have proven that using the LSTM network alone to predict the number of infected cases is not an effective method. Considering that the prevention and control measures and people's awareness of epidemic prevention are closely related to the spread of the virus, this article uses the NLP technology to extract semantic features from news reports related to prevention and control measures and people's attitudes toward the epidemic. These features are then used in the LSTM network. The number of infections is predicted by revising the infection rate in the traditional epidemic model. This method maintains the long-term trend of infectious disease models and updates the infection rate through the usage of news information to improve the accuracy of epidemic prediction.

We collect news information related to the epidemic situation in China. From this information, text data related to prevention and control measures are extracted. The extracted titles and profiles are converted into feature vectors by using a pretrained NLP model. We also extract the features in news information through NLP and combine the LSTM network to update the deviation of the infection rate in the ISI model and achieve an accurate prediction of the number of infections, which is shown in Fig. 3.

Fig. 3.
Prediction model based on the infection rate and NLP features (MLP: multilayer perceptron, NLP: natural language processing, LSTM: long short-term memory network, and CDC: centers for disease control).

A. News Feature Extraction

To extract the relevant features of news information, we examine the information related to the COVID-19, sort this information by date, province, and city, and filter out case reports and related foreign news. Feature extraction is only performed on the title and main content of each news text to obtain concise and robust features in practice. For each given news text in Chinese, a pretrained model of the BERT language model (RoBERTa) [34], which was designed by researchers at Facebook AI and the University of Washington, is used to extract text features. This model combines the Chinese Whole Word Masking strategy, uses WordPiece segmentation to divide a complete word into several subwords, and also combines BERT. This model can achieve good feature extraction results with minimal training.

The news titles and main content are separately obtained as the input to prevent overfitting and achieve efficient training, and the last hidden layer of the pretrained model is used to encode the text. Then, we encode 768-D title and 768-D text together to generate a 1536-D NLP feature vector, in which each vector corresponds to a piece of news. The dataset is divided into national and provincial datasets to achieve accurate daily predictions in different cities and regions and the entire country, where the national dataset contains news from all regions, and the provincial dataset contains news from each province. The news is classified by day to ensure the presence of at least one news per day, and the features of all news of the day are averaged as the NLP feature vector.

B. LSTM Network Based on NLP and Infection Rate

Deep neural networks have the capacity to fit complex distributions but tend to overfit without sufficient supervision. As infection rate features are based on the growing percentage of each factor, they are stable across time. However, epidemic models based on the infection rate cannot predict policy changes and emergency conditions nor adjust the prediction with short-term influence. Therefore, we introduce the LSTM network based on NLP features to model the current policy and social media, which is shown in Fig. 4. Then, the short-term flexibility and long-term stability are both ensured.

Fig. 4.
LSTM network based on NLP features.

In the ISI model, we assume that the actual infection rate is $\beta(t)$, and the infection rate that regressed under the exponential function is $\hat{\beta}(t)$. We use the neural network to predict the bias between the actual infection rate and the regressed infection rate. We let the label of day t be $\mathrm{y}(t)=\beta(t)-\hat{\beta}(t)$, which is taken as the bias feature for prediction. Therefore, we can use the LSTM network as a complement to the ISI model.

To take the impact of news and policies into consideration, we combine the NLP features introduced in Section IV-A with the bias features. We use LSTM to encode temporal information and hidden states. We adopt a one-layer perception model (with a fully connected layer and a leaky ReLU activation function) to transform the infection and NLP features into 32-D vectors. This approach ensures that two features provide the same contribution to our network.

Given infection features s_1 and NLP features s_2 , let the weight of the first two perception model be W_1 and W_2 . Let $g(\cdot)$ be the function of convolution and leaky ReLU as follows:

$$\begin{aligned} f_1 =& g\left(s_1; W_1\right)\\ f_2 =& g\left(s_2; W_2\right). \end{aligned} \tag{6}$$

View Source

The processed features are f_1 and f_2 , which are concatenated into a mixed feature f . At every timestamp t , assuming that the given hidden state from timestamp $t-1$ is h_{t-1} , let the mixed feature be $f(t)$. Function lstm includes the LSTM network and the fully connected layer that transforms the hidden state into prediction. The output of the network is $x(t)$ and the new hidden state $h(t)$. Then

$$\left(x\left(t\right), h\left(t\right)\right) = \mathrm{lstm}\left(f\left(t\right), h\left(t-1\right); W_l\right) \tag{7}$$

View Source

where W_l is the weight of the network. We use gradient descent and the Adam optimizer [35] as the optimization method during training. Then, the mean-square error between prediction and label is adopted as the loss function.

SECTION V.

Experimental Results

In this section, the performance of the proposed model is evaluated on the epidemic data, which are sourced in two ways. First, most of the data mainly come from the national and provincial health commissions and include the numbers of people who are infected, suspected, cured, and have died. Second, data for NLP are obtained from dxy.com [36], social media, and news media. We filter foreign news and disease reports first and then classify the media by the dates and relevant provinces.

A. Correlation Analysis of Cumulative Daily Confirmed Cases and Infection Rate

The infection rate of viruses is deemed to be periodic in existing epidemic models [37]. Considering that patients who have been diagnosed are strictly medically isolated and no longer have the conditions to infect others, we assume that the majority of the sources of infection at day t come from the cumulative new daily confirmed cases in the previous k days. An epidemic model based on a retrospective method is used to analyze the epidemic data of Beijing, Shanghai, and Hunan and further explore the dynamic transmission laws of the virus.

Some patients were missed or misdiagnosed in the early stage of COVID-19 due to the lack of medical resources and changes in diagnostic criteria in certain cities. These factors introduced some noise to the epidemic data. To reduce the impact of noise, we select Shanghai, where public health facilities are relatively complete, as the research object and analyze the evolution laws of the infection rate of the COVID-19. We initially select k time scales to model the correlation between infection rate β_1 and cumulative confirmed cases; then, we analyze the infection rate of the confirmed cases at different time intervals according to the prediction results. The experimental results are of great importance for us to infer the evolutionary trend of the epidemic and estimate the laws of virus infection.

The results of the exponential fitting curves of the estimated infection rate in Shanghai against the different values of k are shown in Fig. 5. Moreover, to quantitatively assess the best value of k determining the infection rate, we use the fitted infection rate to estimate the number of the predicted cumulative confirmed cases. The mean absolute error (MAE) curves between the number of actual cumulative confirmed cases and the number of predicted cumulative confirmed cases for Shanghai are shown in Fig. 6. The infection rate of each epidemic model is obtained from the results of the exponential fitting, and the time scale of the data used for the infection rate fitting is from January 23, 2020 to February 18, 2020. As shown in Fig. 5 and the curve of Shanghai in Fig. 6, when k is small ($k = 1$ –3), the distribution of infection rate β_1 does not show apparent regularity, and the prediction result of the number of cumulative confirmed cases has a relatively large error. This finding shows that the new daily confirmed cases of dates near day t have a weak impact on the infection rate. As k gradually increases ($k = 4$ –6), the distribution of infection rate β_1 becomes concentrated, and the estimation error of the epidemic model decreases rapidly. This observation proves that the trend of infection rate β_1 gradually approaches the truth, and the dates with a significant effect on day t are gradually incorporated into the model. When k is greater than 7, the distribution of β_1 no longer changes significantly, but the MAE curve of the epidemic model gradually increases, proving that the trend of β_1 begins to deviate from the truth. This deviation indicates that noisy data have been introduced into the model, that is, the patients at day $t - k$ have been isolated and no longer infect the S group at day t .

Fig. 5.

Fitting curves of infection rate β_1 in Shanghai. (a) $k=1$. (b) $k=2$. (c) $k=3$. (d) $k=4$. (e) $k=5$. (f) $k=6$. (g) $k=7$. (h) $k=8$. (i) $k=9$. (j) $k=10$.

Fig. 6.

MAE curves between the number of actual cumulative confirmed cases and the number of predicted cumulative confirmed cases in Shanghai, Beijing, and Hunan.

We also establish two epidemic models for patients' data in Beijing and Hunan to verify the generality of the above-mentioned laws. The MAE curves of the two regions are also shown in Fig. 6. According to the performance of the epidemic models, the exponential fitting effect of the infection rate in the two regions is similar to an inverted bell curve, which further validates that the impact of the new daily confirmed cases on the infection rate varies at different dates. The new daily confirmed cases in the middle phase during the period from $t-10$ to $t-1$ have a great impact on the infection rate of time t .

The experimental results of the aforesaid provinces and cities reflect the general development trend of the epidemic, but the change in the diagnosis criteria on February 12, 2020 led to a sharp increase in the number of new daily confirmed cases in Wuhan. To resolve this problem, we initially establish an epidemic model based on data from January 23, 2020 to February 11, 2020 in Wuhan and use the exponential function to fit the overall evolution trend of the infection rate. As shown in Fig. 7, the fitting curve of Wuhan's infection rate is similar to that of Shanghai and

Beijing, that is, the trend of the epidemic in Wuhan is also stable. Therefore, using the data balance method described in Section III-C to preprocess the anomaly data points in Wuhan is reasonable.

Fig. 7.

Fitting curves of infection rate β_1 in Wuhan. (a) $k=1$. (b) $k=2$. (c) $k=3$. (d) $k=4$. (e) $k=5$. (f) $k=6$. (g) $k=7$. (h) $k=8$. (i) $k=9$. (j) $k=10$.

B. Analysis of the Influence on the Infection Rate of New Confirmed Cases at Different Time Intervals

Patients in incubation at different infection time intervals have different infection rates [38], [39]. The new daily confirmed cases from day $t-k$ to day $t-1$ may have different influences on the infection rate of the newly confirmed patients on day t . Here, we investigate the influence and time laws of epidemic transmission in the different provinces and cities by using (2).

We begin by analyzing the relationship between the new confirmed cases in the past ten days and the new daily confirmed cases on day t in Beijing, Shanghai, Zhejiang, and Hunan. Similar to the conclusions in the above section, the curve of parameter α is generally similar to a bell curve when the distribution of weights is considered. That is, the new confirmed cases from day $t-8$ to day $t-3$ have a larger contribution to the new confirmed cases on day t , whereas the contribution rates of the new confirmed cases from $t-10$ to $t-9$ and from $t-2$ to $t-1$ are smaller, as shown in Fig. 8(a).

Fig. 8.

Infection rate of the new confirmed cases from day $t-10$ to $t-1$ to new confirmed cases on day t in the different provinces and cities and the average effect, where "Average" denotes the average contribution of newly confirmed cases from $t-10$ to

$t-1$ to new confirmed cases on day t in four regions: Beijing, Shanghai, Zhejiang, and Hunan. (a) Average. (b) Beijing. (c) Shanghai. (d) Zhejiang. (e) Hunan. (f) Wuhan.

When (2) is used to fit the estimated parameter $\beta_2(t)$, the distribution of α_i shows a trend where the value is small on both sides and large in the middle. Meanwhile, the value of α_i on day $t-10$ is close to 0, indicating that the earlier confirmed cases have little influence on the confirmed cases on day t . Further study reveals that α_i on days $t-8$ to $t-3$ is larger, whereas those on days $t-10$ to $t-9$ and days $t-2$ to $t-1$ are smaller for most provinces and cities. Therefore, the average infection time is about 5.5 days.

To avoid underfitting or overfitting phenomenon analyzed in Section V-A, we balance the parameters via a grouped multiparameter strategy. According to (3), the weights of the dates from $t-8$ to $t-3$ can be set as the same parameter γ_1 , and the weights of the days $t-10$ to $t-9$ and $t-2$ to $t-1$ can be set as the same parameter γ_2 . Then, (3) can be converted to

$$\begin{aligned} I(t) =& I(t-1)+\beta_2(t)\gamma_1\sum_{i=3}^{8}\Delta I(t-i)\\ &+\ \beta_2(t)\gamma_2\left(\sum_{i=1}^{2}\Delta I(t-i)+\sum_{i=9}^{10}\Delta I(t-i)\right) \end{aligned} \tag{8}$$

View Source

where $6\gamma_1+4\gamma_2=1$. According to this equation, we have the results as shown in Fig. 8.

It shows a consistent distribution in the different provinces and cities in Fig. 8. We can see that the values of γ_2 are always less than the values of γ_1 for all the curves in Fig. 8. For Zhejiang and Hunan, γ_2 is close to zero. For other cities, we take the values of γ_2 as a noise and set γ_2 as zero. Finally, we can reformulate (3) as follows:

$$I(t)=I(t-1)+\beta_4(t)\sum_{i=3}^{8}\Delta I(t-i). \tag{9}$$

View Source

C. Prediction of the Cumulative Number of COVID-19 Cases

We verify our model in Wuhan, Beijing, Shanghai, and countrywide. The numbers of preprocessed infections from January 23, 2020 to February 18, 2020 are used as the training data to predict the number of infections from February 19, 2020 to February 24, 2020.

To verify the effectiveness of our model and the influence of government control and public awareness of epidemic prevention, we compare the traditional IS model, the ISI model, the ISI model with the LSTM network, and the ISI model with NLP features and the LSTM network. The NLP features extracted from the current and past news are used in the LSTM network. We compare the daily prediction, MAE, and mean absolute percentage error (MAPE) for Wuhan, Beijing, Shanghai, and countrywide. We round off the prediction results for simplicity, and the compared results are shown in Tables I– IV.

TABLE I Comparison of Actual Confirmed Number and Predicted Number in Wuhan

TABLE II Comparison of Actual Confirmed Number and Predicted Number in Beijing

TABLE III Comparison of Actual Confirmed Number and Predicted Number in Shanghai

TABLE IV Comparison of Actual Confirmed Number and Predicted Number at the Countrywide Scale

Our model makes decent predictions for the three typical cities, as described in Fig. 9. Our ISI model makes a remarkable improvement of the traditional SI model. Compared with the ISI model, the LSTM network is not consistently improved, which is unstable. The ISI+NLP+LSTM achieves a more precise prediction than the other models. This finding shows that NLP features provide extra information and guidance for disease prediction.

Fig. 9.
Comparison of actual confirmed number and predicted number in three typical cities and at the countrywide scale. (a) Wuhan. (b) Beijing. (c) Shanghai. (d) Countrywide.

In summary, based on the ISI model, the hybrid AI model for predicting the COVID-19 proposed in this article is embedded with the NLP module, which introduced the important information of the great efforts led by the central government and local governments as well as the massive support participation from the public into the prediction calculation process, so that the prediction results are more in line with the actual epidemic development trend.

D. Basic Reproduction Number R_0

Basic reproduction number R_0 is a widely applied epidemiologic metric to describe the transmissibility of an infectious patient. In this article, the basic reproduction number $R_0(t)$ is defined as the average number of secondary cases that one confirmed case at time t would produce in a completely susceptible population. According to (9), it is formulated as follows:

$$I(t+j) = I(t+j-1) + \beta_4(t+j)\sum_{i=3}^{8}\Delta I(t+j-i). \tag{10}$$

View Source

According to the above equation, the secondary cases infected by the new daily confirmed cases at time t consist of $\beta_4(t+3)\Delta I(t)$, $\beta_4(t+4)\Delta I(t)$, $\ldots$, $\beta_4(t+8)\Delta I(t)$. Thus, the basic reproduction number at time t can be calculated as

$$R_0\left(t\right) = \frac{\sum_{i=3}^{8}\left[\beta_4\left(t+i\right)\Delta I\left(t\right)\right]}{\Delta I\left(t\right)} = \sum_{i=3}^{8}\beta_4\left(t+i\right). \tag{11}$$

View Source

We analyze the evolutionary trends of the basic reproduction number R_0 in Beijing, Shanghai, Zhejiang, Hunan, and Wuhan, as shown in Fig. 10, from which we can see that the values of R_0 for all regions gradually decrease with the implementation of prevention and control measures.

Fig. 10.

Curves of the basic reproduction number R_0 for different provinces and cities in China.

The Wuhan area was locked down on January 23, 2020, which was a crucial time point of the COVID-19 epidemic. To analyze the impact of the lockdown of the city on R_0 , we analyze more values of R_0 for Wuhan. As shown in Fig. 10, the R_0 curve in Wuhan peaked on January 24, 2020 then dropped rapidly, indicating that locking down the city played an essential role in curbing the spread of the COVID-19. With the proposed hybrid AI model, we also make a prediction about the cumulative confirmed cases in Wuhan, and all of China, the data used for prediction were collected from January 23, 2020 to February 18, 2020. The prediction curves of the cumulative confirmed cases are shown in Fig. 11, from which we can see that the number of cumulative confirmed cases till the end of March would be 48247 for Wuhan. However, if Wuhan was locked down on January 27, 2020, with a delay of four days of the actual time, the number would increase to 102769.

Fig. 11.
Prediction curves of the cumulative confirmed cases in (a) Wuhan and at the (b) countrywide scale.

SECTION VI.
Conclusion

This article, which aims to predict the trend of the COVID-19, discovered that new daily confirmed cases at different time intervals have different contributions to susceptible infections. The impact of confirmed cases in the past several days before time t on the new daily confirmed cases at time t is analyzed. On this basis, we propose a grouped multiparameter strategy that sets the infection rates of the confirmed cases in the past into different groups by time. Then, we derive the proposed ISI model with multiple parameters. This article uses NLP technology to analyze and extract related news information, such as epidemic control measures and residents' awareness of epidemic prevention, which are then encoded into semantic features. Then, these features are fed to the LSTM network to update the infection rate given by the ISI model.

In summary, based on the ISI model, the hybrid AI model for predicting the COVID-19 proposed in this article is embedded with the NLP module, which introduced important information led by the great efforts of the central government and local governments as well as the massive support participation from the public into the prediction calculation process. The prediction results of the model are highly consistent with actual epidemic cases, which proves that the proposed hybrid model can more accurately analyze the transmission law and development trend of the virus compared with previous models and that language information processing of related news can help improve the accuracy of the prediction model. In addition, we provide an effective method for the prediction of the transmission law and development trend of public health events in the future. This article also shows that the openness, transparency, and efficiency of releasing data are very important for establishing a modern epidemic prevention system.

ACKNOWLEDGMENT

The authors would like to thank Dr. P. Hu of the School of Management, Xi'an Jiaotong University, for her discussion and constructive comments for our research framework.

Authors

Figures

References

Citations

Keywords

Metrics

More Like This

Health Education Based on Natural Language Processing(NLP) for Infectious Disease Outbreak
2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)
Published: 2021

Natural Language Processing Methods to Extract Lifestyle Exposures for Alzheimer’s Disease from Clinical Notes
2020 IEEE International Conference on Healthcare Informatics (ICHI)
Published: 2020

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

» Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.