Data Science and Management 3 (2021) 13–21

Contents lists available at ScienceDirect

Data Science and Management

journal homepage: www.keaipublishing.com/en/journals/data-science-and-management

Research article

Improving Google Flu Trends for COVID-19 estimates using Weibo posts

Shuhui Guo [a], Fan Fang [a], Tao Zhou [b], Wei Zhang [c], Qiang Guo [d], Rui Zeng [c,e,*], Xiaohong Chen [f,g,**], Jianguo Liu [h,***], Xin Lu [a,****]

[a] College of Systems Engineering, National University of Defense Technology, Changsha, 410073, China
[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu, 611713, China
[c] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, 610041, China
[d] Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai, 200093, China
[e] MD Department of Cardiology, West China Hospital, Sichuan University, Chengdu, 610041, China
[f] School of Business, Central South University, Changsha, 410083, China
[g] Institute of Big Data and Internet Innovations, Hunan University of Technology and Business, Changsha, 410205, China
[h] Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

Keywords:
COVID-19
Epidemic estimates
Weibo
Google flu trends
Genetic algorithm

ABSTRACT

While incomplete non-medical data has been integrated into prediction models for epidemics, the accuracy and the generalizability of the data are difficult to guarantee. To comprehensively evaluate the ability and applicability of using social media data to predict the development of COVID-19, a new confirmed case prediction algorithm improving the Google Flu Trends algorithm is established, called Weibo COVID-19 Trends (WCT), based on the post dataset generated by all users in Wuhan on Sina Weibo. A genetic algorithm is designed to select the keyword set for filtering COVID-19 related posts. WCT can constantly outperform the highest average test score in the training set between daily new confirmed case counts and the prediction results. It remains to produce the best prediction results among other algorithms when the number of forecast days increases from one to eight days with the highest correlation score from 0.98 ($p < 0.01$) to 0.86 ($p < 0.01$) during all analysis period. Additionally, WCT effectively improves the Google Flu Trends algorithm's shortcoming of overestimating the epidemic peak value. This study offers a highly adaptive approach for feature engineering of third-party data in epidemic prediction, providing useful insights for the prediction of newly emerging infectious diseases at an early stage.

1. Introduction

Since the outbreak of COVID-19 (formally known as 2019-nCoV) in December 2019 in Wuhan, Hubei Province, China (Shen et al., 2020), the pandemic has become a major threat to the whole world. By May 30, 2021, the virus had affected more than 169 million people and caused the deaths of 3.5 million in more than 190 countries and regions worldwide (JHU, 2021). Although many measures have been taken to cope with the health emergency of national concern, such as social distancing measures, locking down measures, imposing quarantines, universities, and business closures (Tison et al., 2020), monitoring the dynamics of the epidemic and preventing its spread poses a huge challenge in practice due to the limited capacity of conventional disease surveillance systems. Studies have shown that publicly available data can play a crucial role in tracking the spread of epidemic disease as complements for conventional public health surveillance (Gundecha and Liu, 2012; Samaras et al., 2020). Non-medical data generated from various sources (Aiello et al., 2020; Kirian and Weintraub, 2010; Ram et al., 2015), has been widely

* Corresponding author. West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, 610041, China.
** Corresponding author. School of Business, Central South University, Changsha, 410083, China.
*** Corresponding author. Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai 200433, China.
**** Corresponding author. College of Systems Engineering, National University of Defense Technology, Changsha 410073, China.
E-mail addresses: zengrui_0524@126.com (R. Zeng), csu_cxh@163.com (X. Chen), liujg004@ustc.edu.cn (J. Liu), xin_lyu@sina.com (X. Lu).

Publishing services by Elsevier on behalf of KeAi

https://doi.org/10.1016/j.dsm.2021.07.001
Received 7 June 2021; Received in revised form 30 June 2021; Accepted 3 July 2021
Available online 15 July 2021

used to estimate disease incidences and to detect disease outbreaks before clinically confirmed data is available (Charles-Smith et al., 2015; Dai et al., 2021; Lu et al., 2021). Social media data collected from Facebook (Gittelman et al., 2015; Strekalova, 2016), YouTube (Basch et al., 2015; Nerghes et al., 2018), Instagram (Guidry et al., 2017; Seltzer et al., 2017), and Internet search queries (Ginsberg et al., 2009; Zhao et al., 2018) are also used to predict diseases for public health concerns. For example, Twitter data is widely used for early warning and outbreak detection, such as to predict syphilis (Young et al., 2018), swine flu (Kostkova et al., 2014), flu (Chen et al., 2014), and Ebola (Yom-Tov, 2015).

The representative work was made by the Google research and development team, who developed the Google Flu Trends (GFT) algorithm based on the high correlation between the number of certain queries in the Google search platform and influenza-like activity level (Ginsberg et al., 2009). They accurately estimated the level of influenza activity in near-real time without knowing the development stage and transmission mechanism of the disease. Since then, many researchers are inspired to track epidemics with social media data (Araujo et al., 2017; Huang et al., 2013; Signorini et al., 2011). As for the unprecedented pandemic COVID-19, some researchers also applied social media and Internet data to monitor and estimate the development of the epidemic (Ayyoubzadeh et al., 2020; Li et al., 2020; Qin et al., 2020). However, many of these studies used only sampled, incomplete data, so the integrity of the dataset and the accuracy of the prediction models are both difficult to guarantee, and there is still a lack of a general prediction framework that can accurately predict the course of COVID-19 using social media data.

To detect and predict the development of COVID-19 using publicly available social media data, this paper applied the daily new confirmed COVID-19 case counts in Wuhan reported by its Health Commission, and a complete dataset of user posts from Sina Weibo (Weibo, 2020), the Twitter-like microblog platform in China, to propose a new confirmed case prediction algorithm named Weibo COVID-19 Trends (WCT) based on the GFT algorithm. WCT can effectively predict the daily new confirmed case counts before the official report is released. This paper also provided a general prediction framework that can be easily extended to predict other diseases or public emergencies using accessible third-party data. This study provides a promising approach for forecasting newly emerging infectious diseases at an early stage when most epidemiological characteristics are unknown. Table 1 shows the nomenclatures used in each processing of this paper.

The main contributions of this paper are summarized as follows:

1. A new confirmed case prediction algorithm is developed based on GFT to predict the development of COVID-19.
2. A genetic algorithm is designed to select a keyword set to filter Weibo posts related to COVID-19.
3. A highly adaptive framework for feature engineering which allows third parties to utilize the data for epidemic predictions is proposed.

Table 1
The nomenclatures used in this paper.

Term	Meaning
GFT	Google Flu Trends algorithm
WCT	The proposed new confirmed case prediction algorithm named Weibo COVID-19 Trends
GCA	The greedy combination algorithm in GFT
LR	Log-odds linear regression model in GFT
GA	Genetic algorithm
LSTM	Long Short-Term Memory regression model
R	Pearson correlation score
KS	Keyword set
MKS	The most epidemic-relevant KS
N	Size of KS
M	Group size of KS in GA
MG	The maximum iteration time in GA
D	Duration of the training data
g	Lag for prediction

The rest of the paper is organized as follows. Section 2 reviews the GFT algorithm and its updated versions. Section 3 mainly describes the framework for the proposed COVID-19 prediction algorithm (i.e., WCT), in which a genetic algorithm is implemented to improve related keyword set selection. Section 4 presents the estimated results of WCT with a comparison with other algorithms including GFT. Finally, Section 5 summarizes the findings and limitations of this study.

2. Literature review

2.1. The initial version of GFT

Google Flu Trends (GFT) is a short-term forecasting tool for weekly influenza activity as an auxiliary method of influenza surveillance (CDC, 2020). It was launched in 2008 with satisfying forecast precision at that time and was further applied to influenza surveillance and early warning systems in many countries (Butler, 2013). Although Google had improved the details of the algorithm many times in the process of GFT application, due to the impact of a sudden increase in influenza-like illness (ILI) related queries and other factors (Kandula and Shaman, 2019; Lazer et al., 2014b). The problem of inaccurate prediction of the algorithm has never been solved completely. Finally, Google shut down the GFT flu prediction function in 2015 (GFT, 2015).

The most well-known GFT algorithm is its initial version. With input on the fraction of certain ILI-related search queries from Google and the percentages of ILI physician visits from the US Centers for Disease Control and Prevention (CDC), the GFT algorithm trains a log-odds linear regression model (LR) to estimate ILI incidence. LR uses the log-odds of an ILI physician visit and the log-odds of an ILI-related search query to realizes regression prediction:

$$\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon \tag{1}$$

where $\text{logit}(p) = \ln(p/(1-p))$, $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t (i.e., the sum of each query fraction in the selected ILI-related search queries set), α is the multiplicative coefficient, and ε is the error term.

Firstly, the model is trained by each of the 50 million candidate common queries separately. It outputs the prediction result of ILI physician visits and the Pearson correlation score between the estimates and the CDC ILI data. Then the aggregated top-scoring queries are used to train the model and the best fit (when the number of keywords $n = 45$) is selected automatically. The selection of queries from the best fit is called "the greedy combination algorithm" (GCA). Finally, the selected queries are used to train the model and predict the ILI physician visits. This approach has successfully estimated the level of weekly influenza activity in the United States from 2007 to 2008 with a mean correlation score of 0.97 and 1–2 weeks ahead of the reports published by CDC. It offers the opportunity to use search queries to detect influenza epidemics and inspires researchers to explore the application of social media data in public health surveillance (Cui et al., 2015; Schmidt, 2012).

2.2. Updated versions and developments

Google officially launched GFT (GFT 1.0) in November 2008, and subsequently gained a wide range of popularity. However, in the first wave of influenza A (H1N1) epidemic, that is, from April to August 2009, the predicted incidence of H1N1 was badly lower than the ILI activity reported by CDC (Butler, 2013). Therefore, Google upgraded GFT for the first time and developed the second version GFT 2.0 (Cook et al., 2011).

GFT 2.0 adjusted the number and category of selected search queries, referring to the ILI monitoring data during the first wave of H1N1 epidemic (March 29 to September 13, 2009). It increased the search

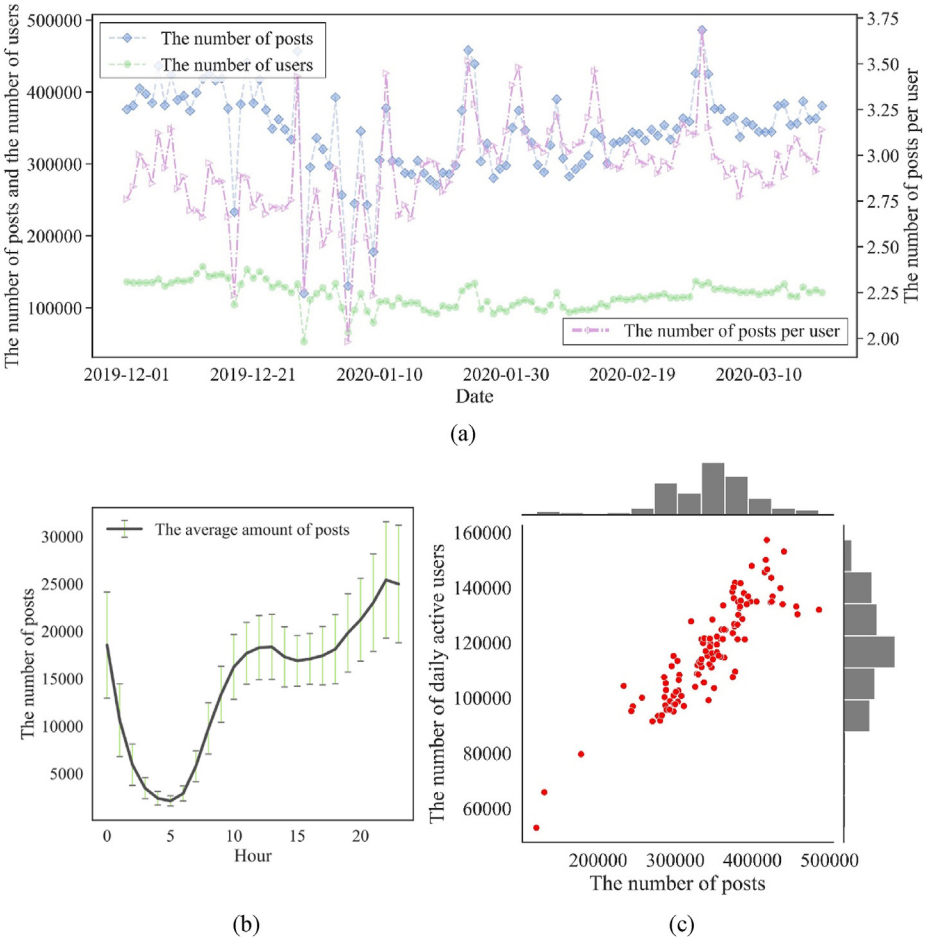

Fig. 1. The statistical description of Sina Weibo dataset. (a) The number of daily posts, users, and posts per user. (b) The average number and standard deviation of posts in the hour 0 to 23 during the statistic period. (c) The correlation between the number of posts and users.

query terms and deleted search queries that were not directly related to influenza, which significantly improved the performance of GFT 2.0. Since its launch in September 2009, its prediction result had been very similar to the ILI activity in the United States until 2012. In the influenza epidemic season of 2012–2013, GFT 2.0 greatly overestimated the influenza epidemic with almost twice the result of CDC monitoring (Butler, 2013). This overestimation led to the second upgrade of GFT (Copeland et al., 2013).

GFT 3.0 was officially launched in October 2013, it made two changes based on GFT 2.0, that is, weakening the impact of abnormal media hot spots and using elastic net to predict ILI (previously based on linear regression). Compared with GFT 2.0, GFT 3.0 significantly reduced the peak amount of its predicted ILI in the 2012–2013 flu season. However, its predicted result was still slightly higher than that of CDC in the United States, and in the 31 weeks after the implementation of GFT 3.0, the prediction result was higher in 23 weeks (Lazer et al., 2014a).

The last upgrade of GFT took place in August 2014 (Lampos et al., 2015). GFT 4.0 expanded the GFT 3.0 model by incorporating the queries selected by the Elastic Net into a non-linear regression framework, based on a composite Gaussian Process. It also injected the ILI activity data as prior knowledge about the disease into the model. The bias of GFT prediction was significantly reduced. GFT 4.0 was used until August 2015, when Google shut down the GFT prediction service.

Because of the important role of ILI surveillance in public health, many researchers are still committed to improving the predictive performance of ILI. Such as correcting the limitations of the GFT algorithm process, updating or adding the training data source of the prediction model, and proposing new prediction algorithms based on GFT. Kandula et al. proposed a corrected GFT algorithm, which uses the estimated value of the original GFT algorithm as new data for training the ILI prediction model, reducing the total prediction error by 44% (Kandula and Shaman, 2019). This algorithm considers the problem that the ILI data provided by CDC is not timely and incomplete when the GFT algorithm is proposed. It uses complete ILI data and GFT estimates to train the prediction model and replaces LR with an autoregressive integrated moving average (ARIMA) model. The algorithm greatly improves the prediction accuracy and proves the validity and practicability of the GFT prediction results. Similarly, other studies (Dugas et al., 2013; Preis and Moat, 2014; Santillana et al., 2015; Wagner et al., 2018) also found that replacing LR with other non-linear regression models and combining new data sources, including search queries, social media, and traditional data sources, into the prediction model can significantly improve the accuracy of ILI prediction.

3. Data and method

3.1. Data description

Sina Weibo is a popular Chinese microblog platform with millions of users voluntarily sharing their lives and thoughts (Weibo, 2020). The considerable amount of post-data generated by so many users offers the possibility of monitoring and predicting the development of emerging infectious diseases. In this study, all posts made by Weibo users in Wuhan from December 1, 2019, to March 20, 2020, were collected. The dataset

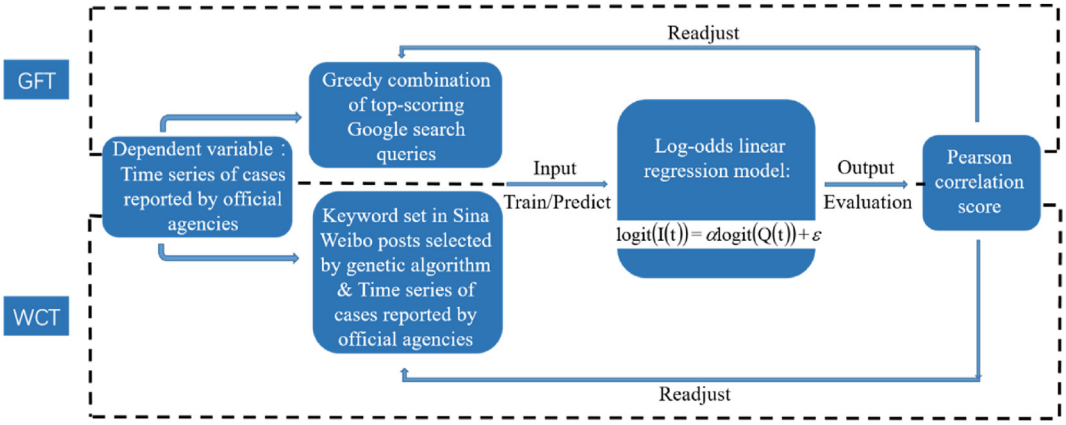

Fig. 2. The basic algorithm flow of WCT and GFT.

spans 111 days and contains the period before the COVID-19 outbreak and its evolution. The dataset contains 38,182,972 posts published publicly by 2,239,450 unique users. Each record of post data contains the post's content, type (whether the post was original or forwarded), time, user nickname, and corresponding encryption ID. If the post was forwarded, the post data contained the original post content (otherwise, it was blank), original time posted, original user nickname, and ID. During the data collection period, the mean number of daily unique users was over 117,000, and they generated more than 343,000 posts every day. On average, each user generated 2.9 posts per day.

Fig. 1a summarizes the series of daily quantity of statistical indicators. The number of posts fluctuated greatly, with a peak of 486,073 posts on March 1, 2020, and the fewest posts (119,886) occurring on December 29, 2019. The number of unique daily users and posts per user remains relatively stable around 117 million and 2.93, respectively. Fig. 1b shows the number of posts from hour 0 to hour 23 of each day. It is obvious that the number of posts decreases first and then increases from hour 0 to hour 23. The minimum value appears at about hour 5, with an average of 2,085 posts, and the maximum value appears near hour 22, with an

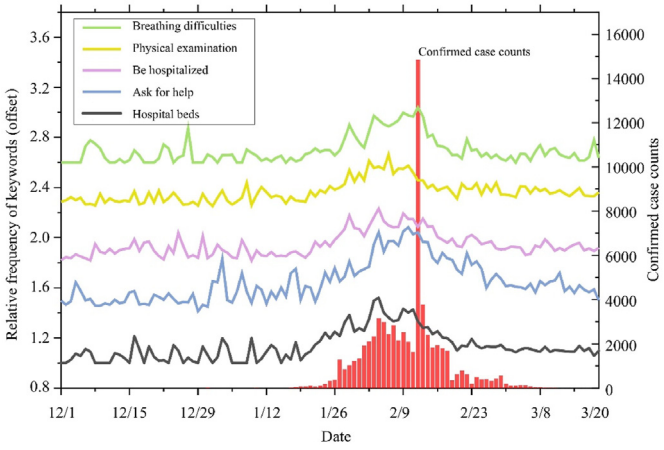

Fig. 4. Evolution of the relative frequency of five most related keywords and the daily case counts.

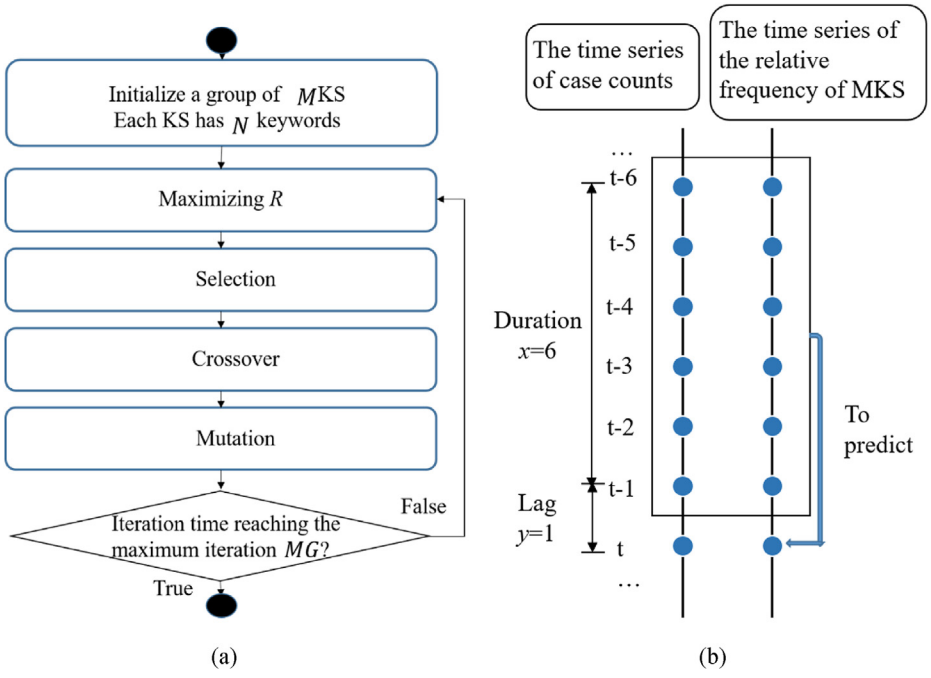

Fig. 3. The MKS selection and case count prediction algorithm processes. (a) MKS selection using GA. (b) The prediction process.

average of 25,395 posts. The number of posts is highly correlated with the number of daily active users (see Fig. 1c), and the Pearson correlation score is 0.89 ($p < 0.01$).

3.2. The framework of weibo COVID-19 trends (WCT)

Inspired by the high correlation score between the relative frequency of the certain keyword in Weibo posts and daily new confirmed case counts of COVID-19 (see Fig. 4a in Section 4), a new confirmed case prediction algorithm named Weibo COVID-19 Trends (WCT) based on GFT is proposed. The basic algorithm process of WCT and its comparison with GFT are shown in Fig. 2. Both of the two algorithms are trying to train a regression model to predict the case counts in which the evaluation indicator is the Pearson correlation score (R) between the prediction results and the real case counts. In WCT, GCA is replaced by the genetic algorithm (GA) (Mitchell, 1998) when selecting the keyword set for the best fit of the prediction model. After comparing the performance of different prediction models, the LR model in GFT is selected as the prediction model in WCT.

3.3. GA for keyword set selection

A prior list of 41 keywords (see Appendix Table A) is compiled firstly to select all posts that contain COVID-19 information, including the pneumonialike epidemic's medical terminology, symptom, and epidemic control measures and organizations. There are 4,761,010 related posts from a total of 38,182, 972 posts from all users (12.47%). Next, the keywords from each post related to the pneumonia-like epidemic were extracted, and a list of 118,572 most commonly used keywords (see Appendix Table B) were produced. The most frequent 2,000 keywords were chosen based on the absolute frequency for the next analysis. The "absolute frequency" of a keyword is the total number of posts containing that keyword since the beginning of the statistical period. Next, the time series of the relative frequency of each commonly used keyword was obtained. The "relative frequency" of a keyword on a certain day refers to the number of all posts containing the keyword on that day divided by the number of unique users on that day.

The relative frequency of a keyword set (KS), i.e., the sum of the relative frequency of each keyword in the selected KS, was used to train the case counts prediction model and then predict the development of the epidemic. The purpose of KS combination and selection is to find the most epidemic relevant keyword set (MKS) from the list of most commonly used keywords in Weibo posts. This paper is aimed to design a selection algorithm to seek the MKS which could obtain the highest R between the prediction results and the real case counts. Viewing the composition of a KS as analogous to an arrangement of chromosomes, GA is used to select the MKS. The fitness function of GA is to maximize R between the prediction results, yielded from the prediction model, and the real case counts. The process of GA is presented as follows:

Step 1 KS initialization. The initial KS group is formed by M KSs, with each KS containing N keywords. Each KS is scored according to the fitness function to maximize R.

Step 2 KS update. The new KS is formed through crossover, mutation, and combination of keywords in KS. Each iteration of the algorithm will choose M better KSs based on R for the next generation and the iteration repeats.

Step 3 Stop criteria. When the maximum iteration time MG is reached or R is high enough, the algorithm will stop and the program will output the MKS.

The flow chart of GA is shown in Fig. 3a. In the implementation process, parameters were set as $M = 25$ and $MG = 100$. Then the respect MKS was obtained with N varying from 1 to 50 while fixing the length of MKS ($N = 1$ *to* 50), separately. To avoid over-fitting, the training period was set as from December 1, 2019, to January 29, 2020, and the test period was set from January 29, 2020, to February 22, 2020. To evaluate the advantages of GA, the MKS obtained by GCA in GFT was also analyzed. The detailed MKS selection results are presented in Section 4.2.

3.4. LR for predicting the number of new confirmed cases

In this section, LR model was applied to predict the number of new confirmed cases using the relative frequency of MKS obtained by GA and a historical case count sequence. The analysis period covers the complete development stage of COVID-19 in Wuhan except February 12 and 13, 2020, due to a change in the criteria for counting diagnoses of the virus. During that period, the number of new confirmed cases increased abnormally. The starting and ending times of the training set and the predicting set are December 1, 2019, to February 21, 2020, and from February 22, 2020, to March 20, 2020, respectively. The case counts series were manually smoothed with a 3-day window length and then used as input data for prediction.

There are also two parameters in the fitting process, the duration (D) of the training data and the lag (g) for prediction. For example, a prediction model trained with $D = 6, g = 1$ is shown in Fig. 3b. In this study, $D = 3$ was set to ensure adequate training data in the training process, and $g = 1$ was set to predict the next day's case counts using all information up to date. All training processes apply three-fold cross validation to reduce overfitting. The training and predicting processes are introduced as follows.

Training process

$$Model_{trained} = FIT_m(C_t, C_{t-g}, C_{t-g-1}, ..., C_{t-g-D+1}, P_{t-g}, P_{t-g-1}, ..., P_{t-g-D+1}) \quad (2)$$

where $Model_{trained}$ is the trained model, C_t and P_t are the case count and number of relative frequency of MKS at time t during the training period, FIT_m is the fitting process by inputting training data $\{C_t, C_{t-g}, C_{t-g-1}, ..., C_{t-g-D+1}, P_{t-g}, P_{t-g-1}, ..., P_{t-g-D+1}\}$ to train $Model_{trained}$. The length of the training window is D and the dimensions of training data is $2D + 1$. The whole training set is $\{C_t, C_{t-g}, C_{t-g-1}, ..., C_{t-g-D+1}, P_{t-g}, P_{t-g-1}, ..., P_{t-g-D+1}\}$ (t increases from 1).

Predicting process

$$C_t = Model_{trained}(C_{t-g-1}, C_{t-g-2}, ..., C_{t-g-D+1}, P_{t-g-1}, P_{t-g-2}, ..., P_{t-g-D+1}) \quad (3)$$

where C_t is the case count at time t during the predicting period. Historical data is input as $\{C_{t-g-1}, C_{t-g-2}, ..., C_{t-g-D+1}, P_{t-g-1}, P_{t-g-2}, ..., P_{t-g-D+1}\}$ into the trained model $Model_{trained}$. Then the prediction result of the case count at time t is output. The length of the predicting window is D and the dimensions of predicting data is $2D$. The whole predicting set is $\{C_{t-g-1}, C_{t-g-2}, ..., C_{t-g-D+1}, P_{t-g-1}, P_{t-g-2}, ..., P_{t-g-D+1}\}$ (t increases from 1).

Previous research has demonstrated that non-linear regression models, such as the Gaussian Processes, Long Short-Term Memory (LSTM), and so on, can achieve great performance in COVID-19 tracking and prediction (Alakus and Turkoglu, 2020; Lampos et al., 2021). The performance of LSTM model was also calculated to be compared with LR model. A 4-layer LSTM model was designed with a dropout rate of 0.15. The loss function was mean square error (MSE) and the optimizer was Adam. The number of training epoch = 100 and batch size = 10. The detailed estimated results are provided in Section 4.3.

4. Results

4.1. Overview of COVID-19 related keywords and case counts

To investigate the relationship between the frequency of COVID-19 related keywords and the number of new confirmed cases per day, the temporal evolution of the keywords with the number of new confirmed COVID-19 cases in Wuhan was analyzed in this section. The direct correlation Pearson score R between the relative frequency of the top 2000 commonly used keywords in Weibo posts and the number of new confirmed cases each day during the whole statistical period was calculated. Most of the correlated keywords are related to the treatment of

Table 2
The keyword combination and performance of MKS selected by four algorithms.

Algorithm	MKS	Length	*R*
GCA&LR	['express', 'Wuhan', 'we', 'pneumonia', 'thanks', 'Zuoyi' (emoji: folded hands), 'virus', 'health commission', 'New Year's Eve Dinner', 'traditional Chinese medicine', 'father', 'Yang Zi' (Chinese star), 'Zhang Wenhong' (Doctor name), 'the Red Cross', 'test kit', 'Boxiao' (Chinese stars: Xiao Zhan and Wang Yibo), '##' (symbol of Weibo topic), 'SARS', 'Super Topic' (Weibo super topics), 'Li Lanjuan' (Doctor name), 'mask', 'coronavirus', 'video', 'suspected case', 'forward', 'Zhang Yixing' (Chinese star), 'you', 'the Red Cross', 'the old people', 'Friend Circle' (Wechat moments), 'investigation', 'hahaha' (laugh), 'husky' (emoji: husky head), 'admission', 'hard work']	35	0.62
GCA&LSTM	['they', 'Hua Chenyu' (Chinese star), 'traditional Chinese medicine', 'Zhao Lei' (personal name), '12', 'video', 'nurse', 'test kit', 'Zhang Yixing' (Chinese star), '14', 'forward', 'ahh' (interjection), 'protective clothing', 'report', 'resume work', ' … ', 'baby', 'medical staff', '17', 'takeaway', 'prevention and control', 'Iran', 'dad', 'doctor', 'Shenshan', 'crying while laughing' (emoji: face with tears of joy), 'CT', 'confirmed', 'N95', 'Han Hong' (Chinese star), 'husky' (emoji: husky head), 'notice', 'Wuhan Union Medical College Hospital', 'Jiang'an District' (a district of Wuhan), 'real', 'Yang Yang' (Chinese star), '9958' (homophony of 'help me')]	37	0.50
GA&LR	['Yang Yang' (Chinese star), 'Li Xian' (Chinese star), 'Han Hong' (Chinese star), 'CT', 'protective clothing', 'anti-epidemic', 'Wuhan City', 'Iran', 'pickup', 'salute', 'Baibuting' (a community of Wuhan), 'Han Hong' (Chinese star), 'Iran', 'Iran', 'Wuhan', 'Iran', 'Wuhan', 'hospital beds', 'Wang Junkai' (Chinese star), 'hospital beds', 'protective clothing', 'the Red Cross', 'hospitalization', '027' (the area code of Wuhan), 'admission', 'community', 'father', 'hospital beds', 'testing', 'Huanggang City', 'Wuhan', 'Iran', 'ahh' (interjection), 'Huanggang City', 'testing', 'Wuhan City', 'Xiao Zhan' (Chinese star), '12', 'admission', 'Wuhan City', 'confession', 'Jiang'an District' (a district of Wuhan), 'Huanggang', 'nucleic acid']	44	0.66
GA&LSTM	['17', 'isolation', 'U.S.', 'mask', 'vaccine', 'satellite TV', 'materials', 'express', 'Korea', 'Zhu Yilong' (Chinese star), 'they', 'Wuhan', 'wild animals', '24', '2020', 'novel coronavirus', 'testing kit', 'the Red Cross', 'CCTV', 'testing kit', 'SARS', 'disappointment', 'hospitalization', '2020', 'Yang Yang' (Chinese star)]	25	0.62

COVID-19 ('hospitalization', 'physical examination', 'patient', and so on), and a few are used to describe symptoms or conditions (such as 'breathing difficulties', 'cough'). The most correlated keywords are 'hospital beds' ($R = 0.84$, $p < 0.01$) and 'Shu Hongbing' ($R = 0.78$, $p < 0.01$). Shu Hongbing is the vice president of Wuhan University and husband of the director of the Wuhan Institute of Virology. The latter was involved in a massive discussion and criticism that it stated that the Chinese herbal remedy Shuanghuanglian can suppress COVID-19. The *R* value, as well as the absolute frequency of the top ten most correlated and uncorrelated keywords, are listed in Appendix Table C.

The evolution of the number of confirmed cases of COVID-19 and the relative frequency of the five most relevant keywords are shown in Fig. 4. It can be seen that the relative frequency of each keyword is very similar to the trend of the number of new confirmed cases, supporting the motivation of tracking COVID-19 with Weibo data. In contrast, the 10 keywords with the weakest correlation ('article', 'new product', '##', 'grandpa Li', 'concert', 'Trump', '19', 'Hubei Economy TV', '2019') were also analyzed. These keywords with low correlation scores have little to do with the symptoms or treatment of COVID-19.

4.2. The R value of the selected MKS

GA and GCA algorithm were both used to select MKS. By setting the length of MKS (N) to vary from 1 to 50 and applying LR and LSTM prediction model ($D = 3$, $g = 1$) into GA and GCA algorithm, the changes in the indicator *R* between the prediction results and the real case counts were compared to evaluate the performance of the MKS selection algorithm. Each prediction model adopted three-fold cross validation and then output the average test scores of the training set as *R*.

The MKSs ($1 \leq N \leq 50$) with the highest *R* selected by each algorithm are presented in Table 2. The original Chinese text for keywords in each MKS are provided in Appendix Table D. Most keywords in MKS obtained by GA or GCA algorithm are medical terms directly related to COVID-19 (such as 'virus', 'isolation', 'CT', 'coronavirus'). It also contains keywords which are not directly related to COVID-19, such as numbers ('14', '17') and personal pronouns ('you'). GA has the feature of retaining the most relevant keywords and automatically outputting MKS with the best performance. The keywords in MKS can be repeated if duplication can make the MKS perform better. It can be found that there are some duplicated keywords in the MKS of GA-related algorithms (see Table 2). This is because the KS with duplicated keywords performs best in the iteration process of GA and becomes MKS. Judging from the correlation between the relative frequency of MKS and the daily case count of COVID-19, the performance of GA and GCA is close, but from the *R* value of the MKS obtained by the two algorithms, GA is better than GCA. The highest test score is obtained by the GA&LR algorithm (WCT) with $R = 0.66$ ($p < 0.01$), which is higher than the test score of GFT (i.e., GCA&LR) of $R = 0.62$ ($p < 0.01$).

In the four combination algorithms, GA&LR (WCT) has the best performance with the average test score $R = 0.65$ ($p < 0.01$), while the average test score of GCA&LSTM is the smallest at $R = 0.43$ ($p < 0.01$). The variation of *R* for MKS with different N is shown in Fig. 5a. Notably, GA-based predictions are much more stable than GCA. For GA&LR and GA&LSTM, the correlation scores vary in a very limited range, 0.60 to 0.66 and 0.55 to 0.62, respectively. However, for GCA-based predictions, the correlation scores experienced unexpected large variations. With GCA&LSTM generating the poorest prediction results, the correlation score of GCA&LR can drop to 0.21 when $N = 50$. In a word, the MKS filtered by GA in terms of predicting daily new confirmed cases is with high agreements to the real data.

In addition, the performances of MKSs filtered by GA and GCA (N from 1 to 50) were compared when the fitness function was to maximize the direct *R* between the relative frequency of the MKS and daily new confirmed case counts. The experimental results further evidenced the superiority of GA in selecting more relevant keyword sets, and it is not sensitive to the length of keywords N (see Figure D8 in Appendix).

4.3. The prediction performance of WCT

In this section, the relative frequency of the selected MKS and daily new confirmed case counts were applied to train prediction models and predict the case counts in the whole analysis period with $D = 3$, $g = 1$. For each prediction result, *R* values between the prediction results and the real case counts in the whole analysis period, the training set, and the predicting set, were calculated as the indicators of performance. Note that different from the three-fold cross validation technique used in the previous analysis, the whole data in the training set were used to construct all models in this section.

The MKSs with the highest *R* selected by GA and GCA were used to train the LR and LSTM model, where the lengths of MKS in GCA&LR, GCA&LSTM, GA&LR, and GA&LSTM are $N = 35$, 37, 44 and 25, respectively (see Table 2). The prediction results show that WCT (referred to GA&LR in Fig. 5b) has a higher prediction accuracy than GFT (referred to GCA&LR in Fig. 5b). The performance of WCT is $R = 0.97$ ($p < 0.01$) during the whole analysis period, all of which are the best among contrast models. While the performance of GFT is $R = 0.96$ ($p <$

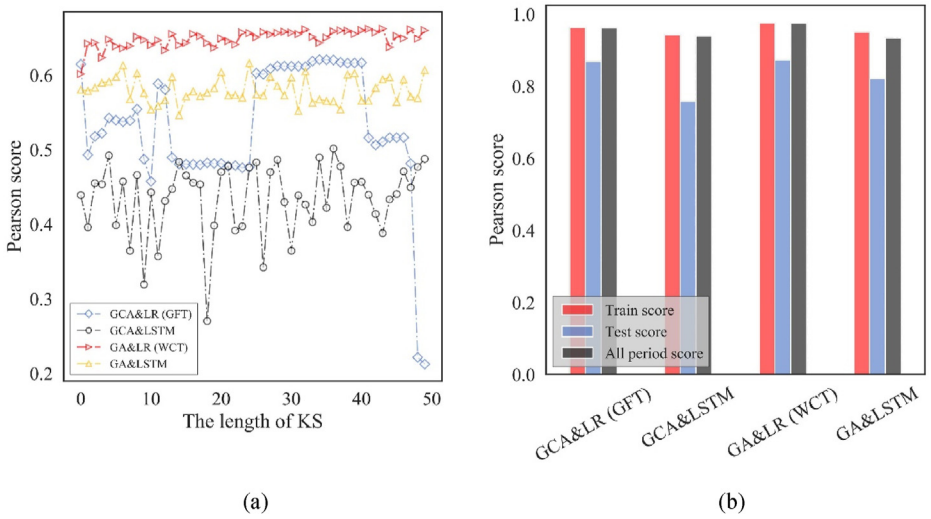

Fig. 5. The MKS and prediction performances. (a) The variation in R of MKSs. (b) The performances of four algorithms.

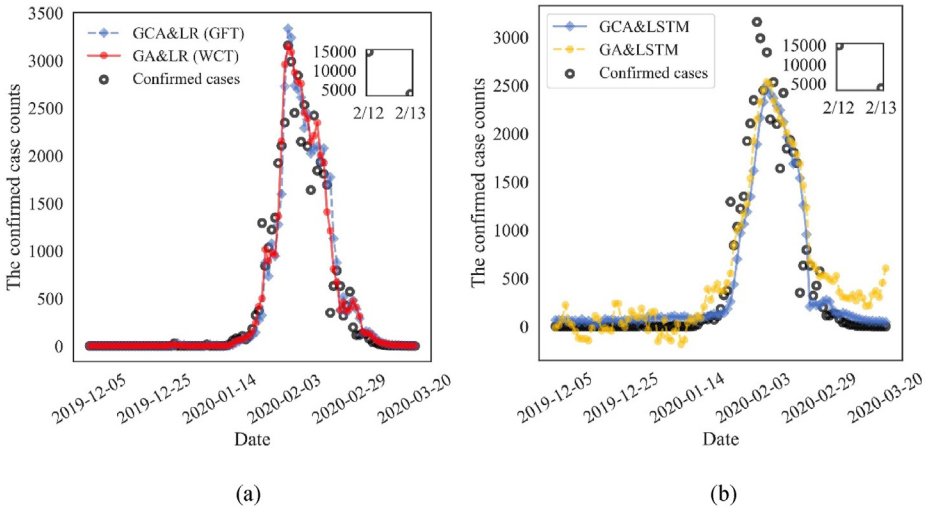

Fig. 6. The daily new confirmed case counts estimates by four algorithms. (a) The estimates by LR-based model. (b) The estimates by LSTM-based model.

0.01). The performance in training set ($R = 0.98$ ($p < 0.01$)) and predicting set ($R = 0.87$ ($p < 0.01$)) of WCT are also the best among the four algorithms.

Compared with GFT, which excessively estimated the daily new confirmed cases during the outbreak period (February 4 to February 5, 2020) over 6–8%, WCT breaks through this limitation and the prediction error is constrained with less than 100 cases (0–3%) (Fig. 6a). The combination of GA and LR effectively overcomes the GFT's shortcoming of over-estimating the epidemic peak value. Besides, in either the training or testing process, WCT constantly outperforms the other algorithms. In contrast, the LSTM model does not perform well in this special task. In both GA&LSTM or GCA&LSTM, the peak number of cases was underestimated by 80% maximumly, and in the late stage of the epidemic, LSTM models overestimated the number of new cases by 10–60% from March 1 to March 22, 2020.

4.4. Sensitivity analysis of WCT

In this section, the performances of the WCT algorithm under different parameter combinations were tested to evaluate the effect of duration of the training data (D) as well as the lag for prediction (g). The parameter D is set to change from 1 to 7, implying that the length of the training window increased from one day to a week before the days to be predicted. The parameter g is set to change from 1 to 15, implying that the algorithm attempts to predict the number of daily new confirmed cases on the gth day in the future. The length of MKS when it produces the best performance in the three-fold cross validation for each algorithm is used in this analysis (see Table 2). Fig. 7 shows the performance of the four algorithms.

The four algorithms all show robustness to the parameter D, especially when g is set in the range of 1–3. When the number of days of historical data used for prediction (D) increases from 1 to 7, the performances of the four algorithms are rather robust, in comparison to the large variation of R in terms of the lag parameter g. Overall, there is a weak tendency of increased performance with larger D, i.e., the prediction model works better when more historical data is included in the training process. When g is small for more recent predictions, the WCT model continues to produce the best result given D is in the range of 2–5. For example, when the algorithm extends the prediction from the next day ($g = 1$) to the second day ($g = 2$) with $D = 3$, the performance of WCT reaches $R = 0.97$ to $R = 0.96$, while the R values of GFT are only 0.96 and 0.93, respectively. When g increases from 10 to 12 with a week's historical data being trained ($D = 7$), the R value of WCT varies in the range of 0.71 to 0.59. On the other hand, GFT only has the R value of 0.59 to 0.51.

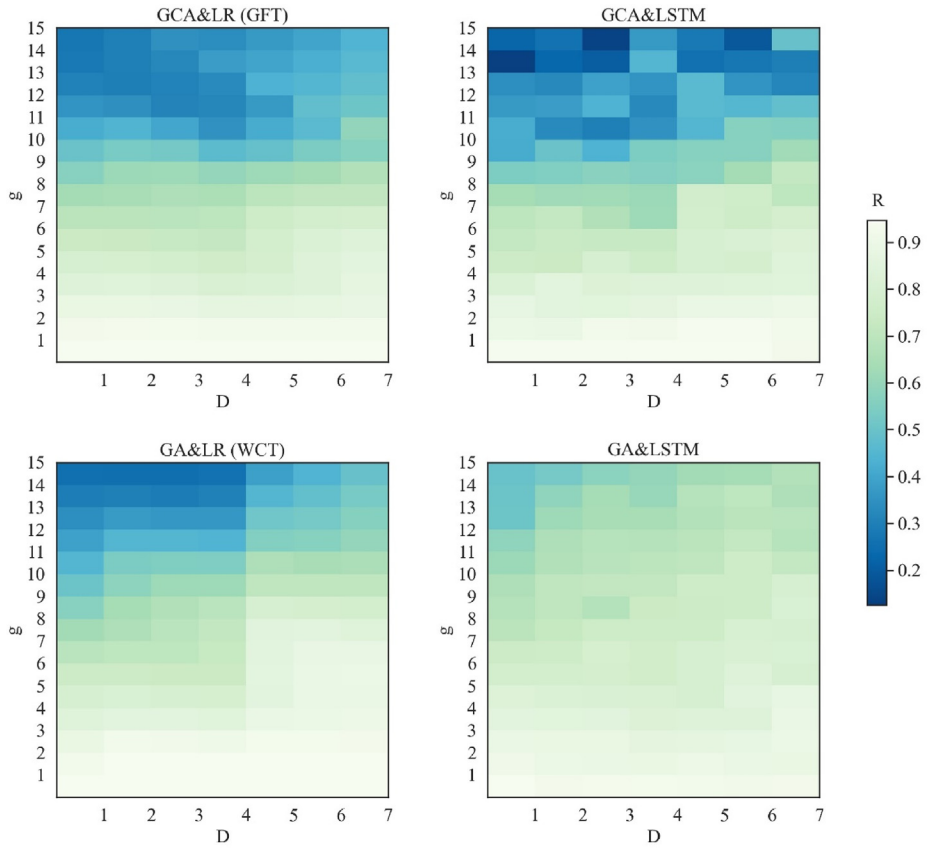

Fig. 7. The prediction performance of the four algorithms with combinations of D and g.

The four algorithms all show sensitivity to the parameter g. As the number of days to predict cases in advance increases, it becomes more difficult for the model to predict the future based on existing data. Compared to the GCA-based algorithms (GFT and GCA&LSTM), GA-based algorithms (WCT and GA&LSTM) show less sensitivity to changes in the g parameter. For example, WCT can still has a great performance as $R = 0.88$ ($D = 6$) when $g = 7$, while the maximum R of GFT is only 0.78 ($D = 7$).

From the comparison of the prediction effect based on the LR model and the LSTM model, the LSTM model is less sensitive to the g parameter and can still maintain a good performance when g increases. WCT remains to produce the best prediction results among other algorithms when the number of forecast days increases from one to eight days with the highest correlation score from 0.98 ($p < 0.01$) to 0.86 ($p < 0.01$). However when g increases to 15, GA&LSTM model can maintain high R as 0.67 ($D = 7$), while WCT is $R = 0.49$, $D = 7$.

Some studies have applied social media dataset to predict new confirmed cases of COVID-19. Qin et al. (2020) used the Baidu search index to predict new confirmed case counts with the performance of $R = 0.99$ for $g = 1$. However, this model is of limited practical value as it was not tested for longer term predictions, on the other hand, the WCT can predict case counts in 1–8 days' future with a high $R = 0.86$-0.98. Lampos et al. designed an unsupervised prediction model using Google Trends data, which can predict newly confirmed case counts with $R = 0.83$-0.85, ahead of official reports in more than 16 days (Lampos et al., 2021). However, this model relies on manual construction of keyword set of Google Trends, which is highly subjective. While WCT utilizes GA to select MKS automatically and heuristically, with little human intervention in the MKS selection process. Ayyoubzadeh et al. (2020) also used Google Trends data to predict newly confirmed case counts in Iran. Comparing linear model and LSTM model, they found that the performance of linear model is better than the LSTM model, which is consistent with the conclusion of this study.

From the above comparison results of sensitivity analysis, it is clear that the WCT method exhibits relatively stronger robustness to the parameters D and g. It produces the highest correlation scores with short future predictions and can maintain relatively more stable performance for longer future estimates.

5. Conclusion and discussion

In this study, an algorithm called WCT is proposed to predict new confirmed cases of COVID-19. Inputting the number of historical case counts and a comprehensive dataset of Sina Weibo posts by Wuhan users, the number of daily new confirmed cases can be accurately predicted by WCT.

This paper applied a genetic algorithm to automatically construct the keyword set and it can consistently outperform the maximum average test score in the training set, higher than that obtained by GCA (0.66 vs. 0.62). The genetic algorithm is more relevant and more stable than GCA in terms of the Pearson correlation score between the prediction results and the real case counts.

The relative frequency of related posts filtered by the selected keyword set is then applied to the LR algorithm and obtained the estimates with a high correlation score of 0.97 ($p < 0.01$) in the whole analysis period one day ahead of the official reports. WCT can accurately predict the development of COVID-19 using only the historical number of cases combined with Weibo post data. Compared with GFT, WCT overcomes the GFT's shortcoming of over-estimating the epidemic peak value.

However, since the development of public emergencies on social media is dynamic, one limitation of the WCT model is that it may need to continuously update the keyword set for future situations with the development of public emergencies, to ensure accurate prediction in the later stage of epidemic or other public emergencies, which makes the application of the method challenging. Compared with the prediction

results of the classical GFT model, considering the influence of noise and other factors, the prediction accuracy of the WCT model in short-term estimates needs to be further improved.

This study offers a promising approach of using Sina Weibo data or other social media data to realize syndromic surveillance-based disease prediction and to increase global awareness of events. It provides a process for mining epidemic development trends from large-scale social media data without too many manual parameters. In the future, the use of WCT can be extended to monitor and track other diseases or public emergencies by inputting social media data.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (91846301, 72025405, 82041020, 11975071, 61773248, 71771152), the Sichuan Science and Technology Plan Project (2020YFS0007), the Hunan Science and Technology Plan Project (2019GK2131, 2020TP1013, 2020JJ4673), the Major Program of National Fund of Philosophy and Social Science of China (18ZDA088, 20ZDA060) and Scientific Research Project of Shanghai Science and Technology Committee (19511102202).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dsm.2021.07.001.

References

Aiello, A.E., Renson, A., Zivich, P.N., 2020. Social media– and internet-based disease surveillance for public health. Annu. Rev. Publ. Health 41 (1), 101–118.
Alakus, T.B., Turkoglu, I., 2020. Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solit. Fractals 140 (Nov.), 110120.
Araujo, M., Mejova, Y., Weber, I., Benevenuto, F., 2017. Using facebook ads audiences for global lifestyle disease surveillance: promises and limitations. Proceedings of the 2017 ACM on Web Science Conference, pp. 253–257.
Ayyoubzadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M., Kalhori, S.R.N., 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR Pub. Health Surveill. 6 (2), e18828.
Basch, C.H., Basch, C.E., Ruggles, K.V., Hammond, R., 2015. Coverage of the ebola virus disease epidemic on youtube. Disaster Med. Public. 9 (5), 531–535.
Butler, D., 2013. When google got flu wrong. Nat. News. 494 (7436), 155. CDC, 2020. CDC's ILI surveillance report. CDC. https://www.cdc.gov/flu/weekly/index.htm. (Accessed 1 June 2021).
Charles-Smith, L.E., Reynolds, T.L., Cameron, M.A., Conway, M., Lau, E.H., Olsen, J.M., Pavlin, J.A., Shigematsu, M., Streichert, L.C., Suda, K.J., et al., 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PloS One 10 (10), e0139701.
Chen, L., Hossain, K.T., Butler, P., Ramakrishnan, N., Prakash, B.A., 2014. Flu gone viral: syndromic surveillance of flu on twitter using temporal topic models. In: IEEE International Conference on Data Mining, pp. 755–760.
Cook, S., Conrad, C., Fowlkes, A.L., Mohebbi, M.H., 2011. Assessing google flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. PloS One 6 (8), e23610.
Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., Stefansen, C., 2013. Google disease trends: an update. In: International Society of Neglected Tropical Diseases.
Cui, X., Yang, N., Wang, Z., Hu, C., Zhu, W., Li, H., Ji, Y., Liu, C., 2015. Chinese social media analysis for disease surveillance. Personal Ubiquitous Comput. 19 (7), 1125–1132.
Dai, B.T., Tan, S.T., Chen, S.R., Cai, M.S., Qin, S., Lu, X., 2021. Measuring the impact of COVID-19 on China's population migration with mobile phone data. Acta Phys. Sin. 70 (6), 068903.
Dugas, A.F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., Rothman, R.E., 2013. Influenza forecasting with google flu trends. PloS One 8 (2) e56176.
GFT, 2015. GFT's Discontinuation Announcement. https://www.google.org/flutrends/about/. (Accessed 1 August 2020).
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.
Gittelman, S., Lange, V., Crawford, C.A.G., Okoro, C.A., Lieb, E., Dhingra, S.S., Trimarchi, E., 2015. A new source of data for public health surveillance: facebook likes. J. Med. Internet Res. 17 (4), e98.
Guidry, J.P., Jin, Y., Orr, C.A., Messner, M., Meganck, S., 2017. Ebola on instagram and twitter: how health organizations address the health crisis in their social media engagement. Publ. Relat. Rev. 43 (3), 477–486.
Gundecha, P., Liu, H., 2012. Mining Social Media: A Brief Introduction, pp. 1–17 chapter (Chapter 1).
Huang, J., Zhao, H., Zhang, J., 2013. Detecting flu transmission by social sensor in China. In: IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, pp. 1242–1247.
JHU, 2021. Coronavirus Resource Center. JHU. https://coronavirus.jhu.edu/. (Accessed 30 May 2021).
Kandula, S., Shaman, J., 2019. Reappraising the utility of google flu trends. PLoS Comput. Biol. 15 (8), e1007258.
Kirian, M.L., Weintraub, J.M., 2010. Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the san francisco bay area. BMC Med. Inf. Decis. Making 10 (1), 1–9.
Kostkova, P., Szomszor, M., St Louis, C., 2014. # swineflu: the use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Trans. Inf. Syst. 5 (2), 1–25.
Lampos, V., Majumder, M.S., Yom-Tov, E., Edelstein, M., Moura, S., Hamada, Y., Rangaka, M.X., McKendry, R.A., Cox, I.J., 2021. Tracking COVID-19 using online search. NPJ Digit. Med. 4 (1), 1–11.
Lampos, V., Miller, A.C., Crossan, S., Stefansen, C., 2015. Advances in nowcasting influenza-like illness rates using search query logs. Sci. Ref. 5 (1), 1–10.
Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014a. Google flu trends still appears sick: an evaluation of the 2013-2014 flu season. Soc. Sci. Electron. Publ. 40 (3), 165–172.
Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014b. The parable of google flu: traps in big data analysis. Science 343 (6176), 1203–1205.
Li, C., Chen, L.J., Chen, X., Zhang, M., Pang, C.P., Chen, H., 2020. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from internet searches and social media data, China, 2020. Euro Surveill. 25 (10), 2000199.
Lu, X., Tan, J., Cao, Z.Q., Xiong, Y.Q., Qin, S., Wang, T., Liu, C.R., Huang, S.Y., Zhang, W., Marczak, L.B., et al., 2021. Mobile phone-based population flow data for the COVID-19 outbreak in mainland China. Health Data Sci. 2021 (1), 9796431.
Mitchell, M., 1998. An Introduction to Genetic Algorithms. MIT Press, Cambridge.
Nerghes, A., Kerkhof, P., Hellsten, I., 2018. Early public responses to the zika-virus on youtube: prevalence of and differences between conspiracy theory and informational videos. In: Proceedings of the 10th ACM Conference on Web Science, pp. 127–134.
Preis, T., Moat, H.S., 2014. Adaptive nowcasting of influenza outbreaks using google searches. Roy. Soc. Open Sci. 1 (2), 140095.
Qin, L., Sun, Q., Wang, Y., Wu, K.F., Chen, M., Shia, B.C., Wu, S.Y., 2020. Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. Int. J. Environ. Res. Publ. Health 17 (7), 2365.
Ram, S., Zhang, W., Williams, M., Pengetnze, Y., 2015. Predicting asthmarelated emergency department visits using big data. IEEE J. Biomed. Health Inform. 19 (4), 1216–1223.
Samaras, L., García-Barriocanal, E., Sicilia, M.A., 2020. Comparing social media and google to detect and predict severe epidemics. Sci. Rep. 10 (1), 1–11.
Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O., Brownstein, J.S., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput. Biol. 11 (10), e1004513.
Schmidt, C.W., 2012. Trending now: using social media to predict and track disease outbreaks. Environ. Health Perspect. 120 (1), A30.
Seltzer, E., Horst-Martz, E., Lu, M., Merchant, R., 2017. Public sentiment and discourse about zika virus on instagram. Publ. Health 150 (Sep.), 170–175.
Shen, C., Chen, A., Luo, C., Zhang, J., Feng, B., Liao, W., 2020. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: observational infoveillance study. J. Med. Internet Res. 22 (5) e19421.
Signorini, A., Segre, A.M., Polgreen, P.M., 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a H1N1 pandemic. PloS One 6 (5), e19467.
Strekalova, Y.A., 2016. Emergent health risks and audience information engagement on social media. Am. J. Infect. Contr. 44 (3), 363–365.
Tison, G.H., Avram, R., Kuhar, P., Abreau, S., Olgin, J.E., 2020. Worldwide effect of COVID-19 on physical activity: a descriptive study. Ann. Intern. Med. 173 (9), 767–770.
Wagner, M., Lampos, V., Cox, I.J., Pebody, R., 2018. The added value of online user-generated content in traditional methods for influenza surveillance. Sci. Rep. 8 (1), 1–9.
Weibo, 2020. Main Page of Sina Weibo. Weibo. Accessed. http://www.weibo.com. (Accessed 1 June 2020).
Yom-Tov, E., 2015. Ebola data from the internet: an opportunity for syndromic surveillance or a news event?. In: Proceedings of the 5th international conference on digital health, pp. 115–119.
Young, S.D., Mercer, N., Weiss, R.E., Torrone, E.A., Aral, S.O., 2018. Using social media as a tool to predict syphilis. Prev. Med. 109 (Apr.), 58–61.
Zhao, Y., Xu, Q., Chen, Y., Tsui, K.L., 2018. Using baidu index to nowcast hand-foot-mouth disease in China: a meta learning approach. BMC Infect. Dis. 18 (1), 1–11.