Original Paper

Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study

Cuihua Shen[1*], PhD; Anfan Chen[2*], PhD; Chen Luo[3], MA; Jingwen Zhang[1,4], PhD; Bo Feng[1], PhD; Wang Liao[1], PhD

[1]Department of Communication, University of California, Davis, Davis, CA, United States

[2]Department of Science Communication and Science Policy, University of Science and Technology of China, Hefei, China

[3]School of Journalism and Communication, Tsinghua University, Beijing, China

[4]Department of Public Health Sciences, University of California, Davis, Davis, CA, United States

[*]these authors contributed equally

Corresponding Author:
Wang Liao, PhD
Department of Communication
University of California, Davis
One Shields Ave
Davis, CA,
United States
Phone: 1 5307520966
Email: wngliao@ucdavis.edu

Abstract

Background: Coronavirus disease (COVID-19) has affected more than 200 countries and territories worldwide. This disease poses an extraordinary challenge for public health systems because screening and surveillance capacity is often severely limited, especially during the beginning of the outbreak; this can fuel the outbreak, as many patients can unknowingly infect other people.

Objective: The aim of this study was to collect and analyze posts related to COVID-19 on Weibo, a popular Twitter-like social media site in China. To our knowledge, this infoveillance study employs the largest, most comprehensive, and most fine-grained social media data to date to predict COVID-19 case counts in mainland China.

Methods: We built a Weibo user pool of 250 million people, approximately half the entire monthly active Weibo user population. Using a comprehensive list of 167 keywords, we retrieved and analyzed around 15 million COVID-19–related posts from our user pool from November 1, 2019 to March 31, 2020. We developed a machine learning classifier to identify "sick posts," in which users report their own or other people's symptoms and diagnoses related to COVID-19. Using officially reported case counts as the outcome, we then estimated the Granger causality of sick posts and other COVID-19 posts on daily case counts. For a subset of geotagged posts (3.10% of all retrieved posts), we also ran separate predictive models for Hubei province, the epicenter of the initial outbreak, and the rest of mainland China.

Results: We found that reports of symptoms and diagnosis of COVID-19 significantly predicted daily case counts up to 14 days ahead of official statistics, whereas other COVID-19 posts did not have similar predictive power. For the subset of geotagged posts, we found that the predictive pattern held true for both Hubei province and the rest of mainland China regardless of the unequal distribution of health care resources and the outbreak timeline.

Conclusions: Public social media data can be usefully harnessed to predict infection cases and inform timely responses. Researchers and disease control agencies should pay close attention to the social media infosphere regarding COVID-19. In addition to monitoring overall search and posting activities, leveraging machine learning approaches and theoretical understanding of information sharing behaviors is a promising approach to identify true disease signals and improve the effectiveness of infoveillance.

(J Med Internet Res 2020;22(5):e19421) doi: 10.2196/19421

KEYWORDS

COVID-19; SARS-CoV-2; novel coronavirus; infectious disease; social media; Weibo; China; disease surveillance; surveillance; infoveillance; infodemiology

Introduction

Since the outbreak of coronavirus disease (COVID-19) in December 2019 in Wuhan, Hubei Province, China [1,2], the novel coronavirus has affected more than 200 countries and territories worldwide. As of May 16, 2020, there were more than 4 million confirmed cases of COVID-19 and over 300,000 deaths [3]. Amid many unknown factors, severe lack of laboratory testing capacity, delays in case reports, variations in local COVID-19 responses, and uncoordinated communication pose tremendous challenges for monitoring the dynamics of the epidemic and developing policies and targeted interventions for resource allocation.

When conventional disease surveillance capacity is limited, publicly available social media and internet data can play a crucial role in uncovering the hidden dynamics of an emerging outbreak [4]. Research in digital disease surveillance, also referred to as infoveillance or infodemiology, has shown great promise in the useful employment of internet data to track the real time development of public attention, sentiment, and health [5-8]. Specifically, data based on internet searches and social media activities can nowcast and forecast disease prevalence as a supplement to conventional surveillance methods for various infectious diseases [5-7,9-14].

One of the best-known examples of digital disease surveillance is Google Flu Trends, which used real time Google search terms to predict clinical incidence rates of influenza with great initial success [13,14]. Data from social media platforms such as Twitter have also been shown to be effective in predicting and tracking various epidemics, such as influenza [10,12] and Zika virus [15], with varying degrees of success. However, digital surveillance data present unique challenges. For example, after its release in 2008, Google Flu Trends became less accurate over time, consistently overestimating flu prevalence during 2011-2013. The prediction error was partially attributed to people's changing search behaviors as well as increased public attention to the epidemic itself, which fueled awareness-related search queries that were not strongly related to disease incidence [7,16]. Compared to aggregated search queries, user-generated social media data have the advantage of being more direct and granular, allowing researchers to mine specific content to reflect actual illness. However, media attention to emerging diseases can fuel social media activities, resulting in a deluge of discussions that dilute true disease signals of actual infection cases; thus, predictions are less accurate [12].

The unprecedented magnitude and transmission speed of COVID-19 brought about massive social media activities as people isolated themselves in their homes to break the infection chain [17]. Massive social media data inevitably contain massive noise (eg, public reactions and awareness of the disease), which can be counterproductive for disease forecasting. A few early infoveillance studies tracked public discussion of COVID-19 and patient characteristics on Weibo, the most popular public social media site in China [18-21]. Two studies suggested that COVID-19–related Weibo posts and search queries can be used to predict disease prevalence [19,22]. However, these studies relied upon coarse-grained social media data and query data based on a few keywords with a short time window at the onset of the outbreak [19,22]. As such, the predictive accuracy and result interpretability of these studies are limited by the same pitfalls of infoveillance studies mentioned above. There are many reasons to search for and discuss COVID-19 on social media, especially because the disease has received substantial media coverage and many countries are under mandatory lockdown. To more accurately predict infection cases and inform a rapid response, it is therefore critical to use granular and specific social media data to identify reliable disease signals (ie, "sick posts" reporting symptoms and diagnosis).

Here, we present an infoveillance effort to collect and analyze COVID-19–related posts on Weibo and to identify specific types of Weibo posts that can predict COVID-19 case counts in mainland China. To our knowledge, this study involves the largest, most comprehensive, and most granular collection of social media data related to COVID-19 in the Chinese language, far exceeding the scale, granularity, and timespan of similar studies [19,22]. We built a Weibo user pool of 250 million people, approximately half the active Weibo user population [23]. Using a comprehensive list of 167 keywords associated with COVID-19, we retrieved around 15 million social media posts from November 1, 2019 to March 31, 2020. With greatly increased data granularity, we developed a supervised machine learning classifier to distinguish "sick posts," which are reports of one's own and other people's symptoms or diagnosis, from other COVID-19 related posts that could dilute disease signals from the data stream. Using the officially reported case counts as the outcome, we compared the predictive power of sick posts versus other COVID-19 posts. We show evidence that sick posts predicted the daily cases reported by the Chinese Center for Disease Control and Prevention (China CDC) up to 14 days in advance, while other COVID-19–related posts had much weaker predictive power. For the subset of geotagged posts, we found that the predictive pattern held true for both Hubei province and the rest of mainland China. Our work demonstrates a viable method to identify disease signals through reports of symptoms or diagnosis rather than relying upon general discussion of COVID-19, making a significant contribution to the infoveillance literature.

Methods

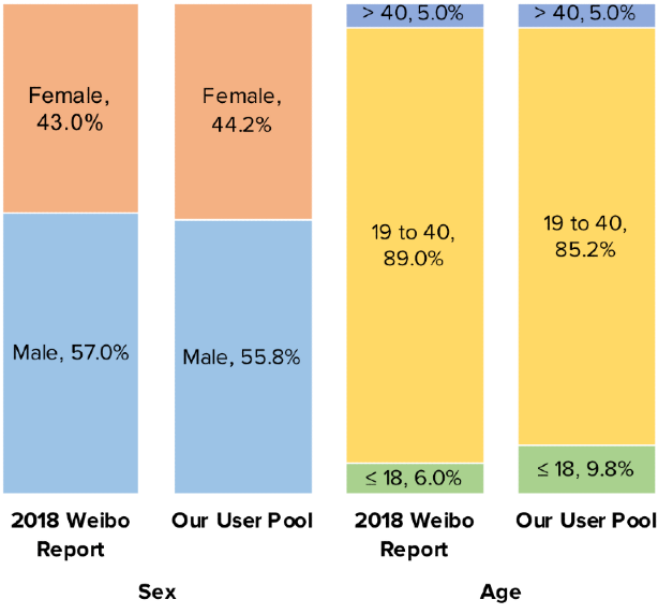
Data Collection

The social media data used in this study were collected from a popular Chinese microblog platform, Weibo, which had over 516 million monthly active users at the end of 2019 [23]. Weibo is very similar to Twitter, access to which is blocked in mainland China. Unlike Twitter, Weibo does not provide large-scale public application programming interface (API) access to its

database. Weibo enables keyword-based advanced searching of posts via its web interface; however, per Weibo policy, the output of these searches is limited to 50 pages (or around 1000 posts). Therefore, large-scale public data access is notoriously difficult.

To bypass these limitations, we employed a Weibo user pool originally built in 2018, which started from 5 million active Weibo users obtained in our previous research unrelated to COVID-19 [24,25]. We then retrieved the initial 5 million users' followers and followees (second degree users), the followers and followees of the second degree users (third degree users), etc., until no new users were found. This snowball process resulted in a pool of 250 million users (with bots filtered out), which represents approximately 48.4% of all monthly active Weibo users in 2019 [23] and is similar to the 2018 population of Weibo users in terms of self-reported sex and age distribution [26] (see Figure 1).

Figure 1. Demographic composition of our Weibo user pool compared to that in the 2018 Annual Sina Weibo user report. Age is reported in years.



COVID-19 Posts

Following best practices for content retrieval and analysis [27], we generated a comprehensive list of keywords related to COVID-19 through close observation of Weibo posts every day from late January to March 2020. We then retrieved COVID-19 posts by searching all posts by users in the user pool with 167 keywords covering general terms related to the epidemic, such as coronavirus and pneumonia, as well as specific locations (eg, "Wuhan"), drugs (eg, "remdesivir") and preventive measures (eg, "mask"). For a complete keyword list, see Multimedia Appendix 1, Table A).

After removing duplicates (ie, reposts of original posts), we retained 14,983,647 posts sent between November 1, 2019 (ie, 30 days before the first confirmed cases) and March 31, 2020 (to access the Weibo dataset on COVID-19, see [28]).

A subset of 464,111/14,983,647 of these posts (3.10%) were tagged with geographic information. We distinguished between posts sent within Hubei province (ie, the epicenter; 169,340/14,983,647; 36.49%) and those from elsewhere in mainland China (294,771/14,983,647; 63.51%).

Sick Posts

We conceptually defined "sick posts" as posts that report any symptoms or diagnoses that are likely related to COVID-19 based on published research and news reports from the medical social media site DXY.cn [29]. We collected a broad list of symptoms, including common symptoms such as cough and shortness of breath and uncommon symptoms such as diarrhea. Sick posts can be further categorized into "ingroup sick posts," which we defined as posts that disclose the user's own or immediate family members' symptoms or diagnoses, and "outgroup sick posts," which report symptoms and diagnoses of people not in the user's immediate family. The reason for the a priori categorization is that people tend to have firsthand and more accurate information about their own or immediate family members' medical conditions; meanwhile, they have much less reliable information about people outside of their household, especially during a national lockdown. All posts that were obtained using the 167 keywords but did not fall into these categories were classified as "other COVID-19 posts." We provide an example of an ingroup sick post below (translated and edited for brevity):

> *During the SARS epidemic in 2003, I got pneumonia with symptoms of fever and cough, was suspected of being infected with SARS, and ended up being hospitalized for more than a month. Now we got COVID-19 in 2020 and I started coughing again, which has lasted for more than a month. What a mess <Face Palm> (Posted 10:23 PM, January 29, 2020)*

We also provide an example of an outgroup sick post:

> *One man in another village drank too much. He said he felt sick and had cold symptoms. His brother measured his temperature which turned out to be 38 Celsius. His brother called 120 and sent him to*

hospital. The whole village was shocked and everyone was afraid to go outside. (Posted 10:14 PM, January 29, 2020)

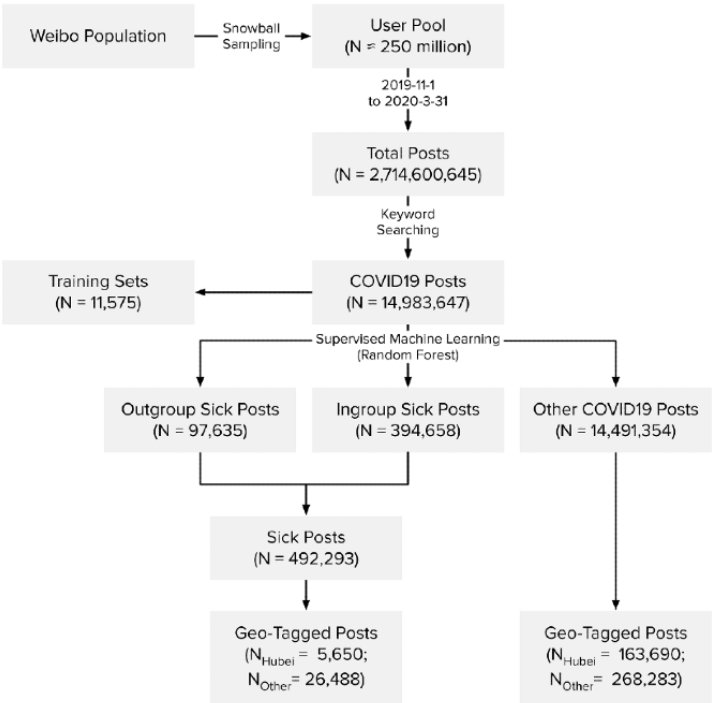
We used supervised machine learning algorithms to identify sick posts from the keyword-retrieved COVID-19 posts. We first sampled 11,575 posts in proportion to the retrieved posts across 5 months of data collection. Next, 11 human judges annotated whether a post was an ingroup sick post, outgroup sick post, or other COVID-19 post. The judges independently annotated a subset of 138 posts and achieved high agreement (Krippendorff α=.945) before they divided and annotated the remaining posts. Then, the annotated posts were used to train machine learning models with various algorithms. Based on the classification performance (see Table 1), we selected the model using the random forest algorithm (F1 score=0.880). The model classified the 14,983,647 COVID-19 posts into 394,658 (2.63%) ingroup sick posts, 97,635 (0.65%) outgroup sick posts, and 14,491,354 (96.71%) other COVID-19 posts. Because of the low number of outgroup sick posts, we combined ingroup and outgroup sick posts in subsequent analyses.

Among the subset of geotagged COVID-19 posts (464,111/14,983,647, 3.10% of all retrieved posts), 5,650 sick posts (1.2%) and 163,690 other COVID-19 posts (35.3%) were tagged in Hubei; meanwhile, 26,488 sick posts (5.7%) and 268,283 other COVID-19 posts (57.8%) were from elsewhere in mainland China. These post counts were then aggregated by days. To control for the day-to-day fluctuations of Weibo posts, we further normalized these numbers against the daily counts of all Weibo posts generated by our user pool. The normalized sick post and other COVID-19 post counts can be interpreted as counts per 1 million posts. Figure 2 summarizes our data collection and classification process.

Table 1. Performance of machine learning models in classifying sick posts.

Model	F1 score	Precision	Accuracy	Recall
Decision tree	0.835	0.840	0.830	0.830
Extra tree	0.785	0.785	0.785	0.785
Extra trees	0.878	0.881	0.885	0.885
K nearest neighbors	0.810	0.819	0.819	0.819
Multilayer perceptron	0.847	0.845	0.851	0.851
Support vector machine	0.877	0.877	0.878	0.878
Random forest	0.880	0.885	0.888	0.888

Figure 2. Weibo data collection and classification procedure.



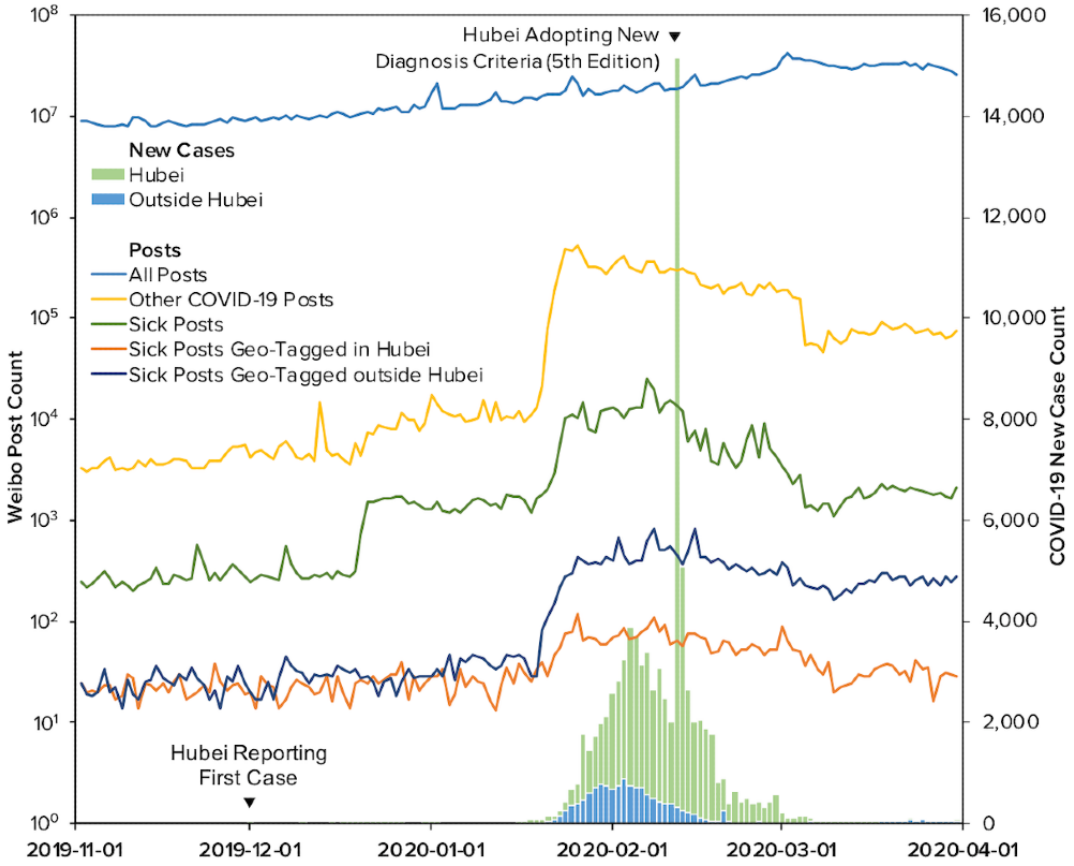
COVID-19 Daily Case Counts

We collected the daily new case counts in mainland China from China CDC on May 8, 2020. China CDC's official website started collating data on January 16, 2020. Earlier counts were obtained from Huang et al [1] and validated against relevant briefings from the National Health Commission. The final case data cover the same period from November 1, 2019, to March

31, 2020, within which the first reported COVID-19 clinical case dates back to December 1, 2019. We also distinguished between cases within and outside Hubei (see Figure 3).

It is noteworthy that China CDC released seven editions of diagnostic criteria throughout the time period covered in this study and thus introduced systematic changes to the case counts. Particularly, on February 12, 2020, Hubei province started to implement the fifth edition of the COVID-19 diagnostic criteria released on February 4, 2020. This led to a temporary surge of new cases [30]. The impact of this incident was controlled for in our analyses, as discussed in the section below. After close comparison of each edition, we concluded that the changes among other editions of the diagnostic criteria were relatively minor, and their release dates did not appear to be associated with abrupt changes in the case counts; therefore, we did not further control for them.

Figure 3. Daily Weibo posts and confirmed COVID-19 cases between November 1, 2019 and March 31, 2020.



Statistical Analysis

We performed Granger causality tests [31] to discover if an increase of sick posts forecasted an increase of new cases, as formulated in the following linear model:

$$\Delta C_t = a_0 + \sum_{i=1}^{m} a_i \, \Delta C_{t-i} + \sum_{j=1}^{m} b_j \, \Delta S_{t-i} + c_1 I_t + \varepsilon_t$$

where Ct is the difference in new case counts at day t from day $t-1$, S_{t-i} is the difference in sick post counts (normalized) at day t from day $t-1$, and I_t is a time-varying binary variable that equals 1 on February 12, 2020, the day on which Hubei adopted the fifth edition of the diagnostic criteria. This binary variable controls for the exogenous pulse of case counts [32]. Since we collected Weibo posts from as early as November 1, 2019, 30 days before the first reported case of COVID-19 on December 1, 2019, we were able to test up to 29 lags of such posts (ie, $m \leq 29$). The model is further explained as follows.

First, difference scores instead of raw new case counts were used because Dickey-Fuller tests for the raw counts could not reject nonstationarity (ie, the presence of a unit root) for lag 3–29 at a 5% confidence level (see Table B in Multimedia Appendix 1). Both stationarity and the inclusion of autoregressive terms are required by Granger causality. In contrast, the Dicky-Fuller tests suggested that the difference scores of the case counts were stationary: nonstationarity was rejected for lag 1–12 at a 1% confidence level and for lag 13–29 at a 5% confidence level (see Table B in Multimedia Appendix 1). The Dickey-Fuller tests reached the same conclusion for the stationarities of the sick post counts and their difference scores (see Table B in Multimedia Appendix 1). We thus also used the difference scores instead of the raw counts to reduce correlations among lag terms of sick post counts. This more clearly identifies their independent effects on case counts. In short, these difference scores can be interpreted as "daily-additional" cases or Weibo posts in addition to the counts from the previous day.

Second, to determine the number of lag terms to include (ie, m in the above formula), we compared model fit statistics while iteratively adding lag terms. The model comparison suggested

that the inclusion of more lags continuously improved the model fit up to the maximum lags (ie, 29; see Table C in Multimedia Appendix 1). However, the parameter estimates did not change qualitatively after including more than 20 lags (see Tables D and E in Multimedia Appendix 1). For parsimony and statistical power, we settled at 20 lags for the following analyses.

Finally, we included a binary variable to control for the change in the diagnostic criteria of COVID-19 on Feb 12, 2020, following the procedure of intervention analysis [33]. Because this change is unlikely to induce permanent changes to case counts, an instant pulse function was applied at the date of the change. We also tested models that allowed the effect to linearly decay in 2, 3, 4, or 5 days; these models fitted the data more poorly than the model with an instant pulse (see Table F in Multimedia Appendix 1).

Results

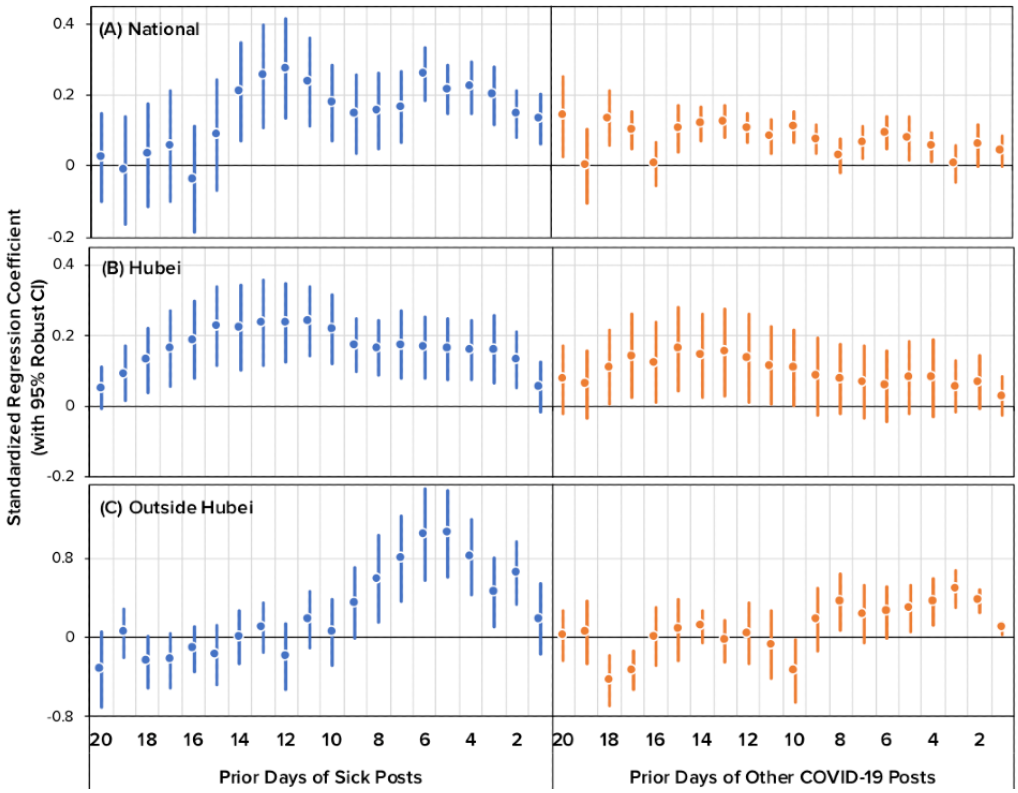
Ordinary least squares regression with robust standard errors was used to estimate the final models. With 20 lag terms in the model, the modeled data include daily-additional new COVID-19 cases from December 1, 2019 to March 31, 2020 and daily-additional counts of sick posts and other COVID-19 posts from November 10, 2019 to March 11, 2020 (N=122).

Figure 4A summarizes the estimates of Granger causality for sick posts predicting new COVID-19 cases with standardized regression coefficients (see Table G in Multimedia Appendix 1 for all estimated parameters). Particularly, one standard deviation of increase in the daily-additional sick posts (1 sick post per 1 million posts) predicted a 0.133 (95% CI 0.065-0.201) to 0.275 (95% CI 0.134-0.416) standard deviation of the increase in the daily-additional new cases 1-14 days in advance. After including the 20 lags of sick posts, the adjusted R^2 value of the model increased by 0.128, suggesting that sick posts could explain an additional 12.8% of the variance of daily-additional new cases beyond the autoregressive terms and intervention effects.

Furthermore, we estimated the relationship between other COVID-19 post counts and daily-additional new cases using the same linear model. Figure 4A further illustrates the standardized estimates. Compared with sick posts, other COVID-19 posts were weaker signals of future case counts, as demonstrated by their smaller standard regression coefficients. This indicates that Weibo posts that discussed some aspect of COVID-19 but did not explicitly report a person's symptoms or diagnosis had lower forecasting power than sick posts.

To corroborate the above results, we tested the Granger causality of sick posts on cases within Hubei and outside Hubei (see Table H in Multimedia Appendix 1). Within Hubei, the results generally agreed with the national pattern mentioned above. Daily-additional sick posts predicted daily-additional new cases in Hubei up to 19 days in advance, as illustrated in Figure 4B. In contrast, other COVID-19 posts had fewer lag terms that could forecast new cases. Outside Hubei, the predictive pattern of sick posts was similar to the national pattern despite a limited time range: sick posts could forecast new cases 2 to 8 days in advance (see Figure 4C).

Figure 4. Standardized estimates of Granger causality for time-lagged, daily-additional Weibo posts (sick posts and other COVID-19 posts) predicting daily-additional cases.



Discussion

Principal Findings

The novel coronavirus causing COVID-19 is a new pathogen in the human reservoir. It poses an extraordinary challenge for public health systems worldwide because screening and diagnostic tests must be developed from scratch. Even when such tests eventually become available, testing capacity is often severely limited; this can fuel the outbreak, as many patients can unknowingly infect other people. Based on approximately 15 million COVID-19-related Weibo posts between November 1, 2019 and March 31, 2020, we developed a supervised machine learning classifier to identify "sick posts," in which a user reports their own or other people's symptoms and diagnosis of COVID-19. Using the officially reported daily case counts as the outcome, our work shows that sick posts significantly predict daily cases up to 14 days ahead of official statistics. This finding confirms prior research that social media data can be usefully applied to nowcasting and forecasting emerging infectious diseases such as COVID-19 [22,34].

One of the greatest challenges of digital disease surveillance is identifying true disease signals, especially when facing the deluge of social media activity that resulted from COVID-19 mitigation measures [12,34-36]. Our finding that sick posts have greater predictive power than other COVID-19 posts shows that not all social media data are equally informative. Specifically, COVID-19 has dramatically disrupted everyday life; due to the pandemic, people are sheltering in place and increasingly communicating with others via social media. As shown in prior work [18] as well as in our data set, the majority of COVID-19–related chatter on Weibo reflected public awareness of COVID-19 rather than actual symptom reports. Most previous studies took rather coarse-grained approaches, relying primarily on either aggregated search query data or social media data retrieved from limited keyword searches [19,22]. In our work, we gathered the largest, most comprehensive, and most granular collection of social media data related to COVID-19 in the Chinese language. More importantly, we demonstrate a viable method to separate valid signals from noise using reports of symptoms and diagnosis, which makes a significant contribution to the literature on digital surveillance.

Another important finding is that while the predictive power of sick posts on daily case counts holds true for both Hubei and non-Hubei regions, the effect sizes vary. Being the epicenter of the outbreak, Hubei province experienced extreme testing shortages during the early stage of the study period. As a result, many Hubei residents turned to social media sites such as Weibo to seek help for testing and medical care. In contrast, social media help-seeking activities were uncommon in other parts of China, where testing and health care resources were much more adequate. Taking these regional variations into account, we still observed predictive signals of sick posts on case counts, suggesting that the predictive power of sick posts was robust against testing delays. Further, the variations in the effect estimates show that the predictive power of social media data may vary across different geographic areas, with different levels of preparedness, and at different stages of the outbreak. Future studies based on longer periods of data monitoring could explore the temporal and spatial variations of COVID-19 social media surveillance efficacy in more depth.

Our work has broad public health implications. The high speed and low cost of social media surveillance can be especially useful in the early stages of the COVID-19 outbreak to inform containment and mitigation efforts when they are most cost-effective. For countries and regions where public health infrastructures do not allow for widespread screening and diagnostic tests, social media disease surveillance provides much-needed information for public health agencies to model the trajectories of the outbreak and to make swift decisions about allocation of resources such as hospital beds, ventilators, and personal protective equipment.

Another advantage of social media surveillance is that it can be performed from a distance. As COVID-19 continues to spread worldwide, countries lacking testing and screening infrastructures will become "dark spots," endangering their own citizens as well as the entire world. It is imperative that international organizations such as the World Health Organization integrate such data into their outbreak forecasting management practices to mobilize and coordinate relief efforts to help combat COVID-19.

Limitations

This study has several limitations. First, Weibo posts were retrieved retrospectively rather than in real time; therefore, deleted or censored posts were absent from our data set. However, we have no reason to believe that deletion or censorship favored "sick posts" in measurable ways. In fact, a recent study on Weibo censorship from December 2019-February 2020 shows that only 1.7/1000 Weibo posts were censored; also, these censored posts generally pertained to the missteps in the government's COVID-19 response, not individual reports of symptoms and diagnoses [37]. Therefore, our results should not be affected by censorship. Second, as some studies suggest [38-40], confirmed COVID-19 case counts published by China CDC may underestimate the actual counts, due in part to limits in testing capacity and the existence of asymptomatic carriers. Still, the data here represent the best-known data of confirmed case counts, and our models rely on trends and changes in these case counts rather than the actual numbers. Third, it is important to acknowledge that sick posts as disease signals are not without noise because Weibo users who reported COVID-19 symptoms were not necessarily clinically diagnosed with COVID-19; Weibo users may not speak the truth; and Weibo users may "overreport" (posting about their symptoms or diagnoses multiple times) or "underreport" (not posting despite their symptoms or diagnoses) for a variety of reasons. Such inaccuracies are inherent in user-generated social media data and widely exist in all infoveillance studies. However, it should be noted that the goal of infoveillance has never been to achieve one-for-one matching between social media posts and clinical cases. Rather, infoveillance approaches strive to mine useful early signals from social media and internet data as a supplement to conventional surveillance efforts. Despite this noise, we still found that sick posts predicted COVID-19 case counts,

indicating the validity of this signal in reflecting disease spread in the population.

Conclusions

The threats of COVID-19 and other infectious diseases are likely to recur in the future. Reports of symptoms and diagnoses on social media during emerging disease outbreaks send invaluable warning signals to the public. Researchers and disease control agencies should pay close attention to the social media infosphere. In addition to monitoring overall search and posting activities, it is crucial to sift through the contents and efficiently separate true signals from noise. Our main findings highlight the importance of using rigorous procedures and understanding information sharing behaviors to obtain quality disease signals. Future studies based on longer periods of data monitoring could explore the time and spatial diffusions of COVID-19 in more depth. A more detailed examination of post contents reporting restraints in information or medical resources will be helpful in developing local outbreak responses.

Acknowledgments

We thank Jingyang Xu, Minwei Ren, Rixia Tang, Zichao Wang, Yongyan Xu, Na Yang, Yalan Jin, Xiuchan Xu, Xinyu Wang, Ruizhi Sun, Wenhui Zhu, Yiwei Li, and Tianyu Zhao for their help with data annotation.

Authors' Contributions

CS, WL, JZ, and BF contributed to the study design. AC collected the Weibo data. WL, CL and AC contributed to the data analysis. WL, CS, CL, and AC contributed to the design and drawing of the figures. All authors contributed to the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables.
[DOCX File , 182 KB-Multimedia Appendix 1]

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb;395(10223):497-506. [doi: 10.1016/s0140-6736(20)30183-5]
2. Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020 Feb 3;579(7798):265-269. [doi: 10.1038/s41586-020-2008-3]
3. World Health Organization. 2020 May 16. Coronavirus disease 2019 (COVID-19) Situation Report 117 URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200516-covid-19-sitrep-117.pdf?sfvrsn=8f562cc_2 [accessed 2020-05-26]
4. Zhang J, Centola D. Social Networks and Health: New Developments in Diffusion, Online and Offline. Annu Rev Sociol 2019 Jul 30;45(1):91-109. [doi: 10.1146/annurev-soc-073117-041421]
5. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
6. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
7. Aiello AE, Renson A, Zivich PN. Social Media- and Internet-Based Disease Surveillance for Public Health. Annu Rev Public Health 2020 Apr 02;41:101-118. [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
8. Barros JM, Duggan J, Rebholz-Schuhmann D. The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
9. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One 2015;10(10):e0139701 [FREE Full text] [doi: 10.1371/journal.pone.0139701] [Medline: 26437454]
10. Cui X, Yang N, Wang Z, Hu C, Zhu W, Li H, et al. Chinese social media analysis for disease surveillance. Pers Ubiquit Comput 2015 Sep 11;19(7):1125-1132. [doi: 10.1007/s00779-015-0877-5]
11. Fung IC, Fu K, Ying Y, Schaible B, Hao Y, Chan C, et al. Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks. Infect Dis Poverty 2013 Dec 20;2(1):31 [FREE Full text] [doi: 10.1186/2049-9957-2-31] [Medline: 24359669]

12. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
13. Klembczyk JJ, Jalalpour M, Levin S, Washington RE, Pines JM, Rothman RE, et al. Google Flu Trends Spatial Variability Validated Against Emergency Department Influenza-Related Visits. J Med Internet Res 2016 Jun 28;18(6):e175 [FREE Full text] [doi: 10.2196/jmir.5585] [Medline: 27354313]
14. Dugas AF, Hsieh Y, Levin SR, Pines JM, Mareiniss DP, Mohareb A, et al. Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. Clin Infect Dis 2012 Feb 15;54(4):463-469 [FREE Full text] [doi: 10.1093/cid/cir883] [Medline: 22230244]
15. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295 [FREE Full text] [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
16. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]
17. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao T, et al. Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. IEEE Trans Comput Soc Syst 2020 Apr;7(2):556-562. [doi: 10.1109/tcss.2020.2980007]
18. Zhu Y, Fu K, Grépin KA, Liang H, Fung IC. Limited Early Warnings and Public Attention to Coronavirus Disease 2019 in China, January-February, 2020: A Longitudinal Cohort of Randomly Sampled Weibo Users. Disaster Med Public Health Prep 2020 Apr 03:1-4 [FREE Full text] [doi: 10.1017/dmp.2020.68] [Medline: 32241328]
19. Li J, Xu Q, Cuomo R, Purushothaman V, Mackey T. Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study. JMIR Public Health Surveill 2020 Apr 21;6(2):e18700 [FREE Full text] [doi: 10.2196/18700] [Medline: 32293582]
20. Zhao Y, Cheng S, Yu X, Xu H. Chinese Public's Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study. J Med Internet Res 2020 May 04;22(5):e18825 [FREE Full text] [doi: 10.2196/18825] [Medline: 32314976]
21. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X, et al. Mining the Characteristics of COVID-19 Patients in China: Analysis of Social Media Posts. J Med Internet Res 2020 May 17;22(5):e19087 [FREE Full text] [doi: 10.2196/19087] [Medline: 32401210]
22. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Euro Surveill 2020 Mar;25(10):1 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.10.2000199] [Medline: 32183935]
23. weibo.com. 2019 Annual Sina Weibo User Report URL: http://ir.weibo.com/node/7726/html
24. Chen Z, Su CC, Chen A. Top-down or Bottom-up? A Network Agenda-setting Study of Chinese Nationalism on Social Media. J Broadcasting Electron Media 2019 Sep 20;63(3):512-533. [doi: 10.1080/08838151.2019.1653104]
25. Li Y, Luo C, Chen A. The evolution of online discussions about GMOs in China over the past decade: Changes, causes and characteristics. Cultures of Science 2020 Jan 20;2(4):311-325. [doi: 10.1177/209660831900200406]
26. weibo.com. 2019. 2018 Annual Sina Weibo User Report. Webpage in Chinese URL: https://data.weibo.com/report/reportDetail?id=433
27. Lacy S, Watson BR, Riffe D, Lovejoy J. Issues and Best Practices in Content Analysis. Journal Mass Commun Q 2015 Sep 28;92(4):791-811. [doi: 10.1177/1077699015607338]
28. Hu Y, Huang H, Chen A, Mao XL. arXiv. 2020 May 21. Weibo-COV: A Large-Scale COVID-19 Social Media Dataset from Weibo URL: https://arxiv.org/abs/2005.09174 [accessed 2020-05-26]
29. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit Health 2020 Apr;2(4):e201-e208 [FREE Full text] [doi: 10.1016/S2589-7500(20)30026-1] [Medline: 32309796]
30. China Center for Disease Control and Prevention. 2020 Feb 12. COVID-19 Situation Report on Feburary 12, 2020. Webpage in Chinese URL: http://www.chinacdc.cn/jkzt/crb/zl/szkb_11803/jszl_11809/202002/t20200213_212624.html [accessed 2020-05-26]
31. Granger CWJ. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 1969 Aug;37(3):424. [doi: 10.2307/1912791]
32. Box GEP, Tiao GC. Intervention Analysis with Applications to Economic and Environmental Problems. J Am Stat Assoc 1975 Mar;70(349):70-79. [doi: 10.1080/01621459.1975.10480264]
33. Box-Steffensmeier J, Freeman J, Hitt M, Pevehouse J. Time Series Analysis for the Social Sciences. Cambridge, UK: Cambridge University Press; 2014.
34. Buckee C. Improving epidemic surveillance and response: big data is dead, long live big data. Lancet Digit Health 2020 May;2(5):e218-e220 [FREE Full text] [doi: 10.1016/s2589-7500(20)30059-5]
35. Hua J, Shaw R. Corona Virus (COVID-19) "Infodemic" and Emerging Issues through a Data Lens: The Case of China. Int J Environ Res Public Health 2020 Mar 30;17(7):2309 [FREE Full text] [doi: 10.3390/ijerph17072309] [Medline: 32235433]

36. Leung GM, Leung K. Crowdsourcing data to mitigate epidemics. Lancet Digit Health 2020 Apr;2(4):e156-e157. [doi: 10.1016/s2589-7500(20)30055-8]
37. Fu K, Zhu Y. Did the world overlook the media's early warning of COVID-19? J Risk Res 2020 Apr 24:1-5. [doi: 10.1080/13669877.2020.1756380]
38. Kucharski A, Russell T, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020 May;20(5):553-558 [FREE Full text] [doi: 10.1016/S1473-3099(20)30144-4]
39. Imai N, Dorigatti I, Cori A, Donnelly C, Riley S, Ferguson N. Imperial College. 2020 Jan 22. Report 2: Estimating the potential total number of novel Coronavirus (2019-nCoV) cases in Wuhan City, China URL: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-update-epidemic-size-22-01-2020.pdf [accessed 2020-05-26]
40. Wu J, Leung K, Leung G. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020 Feb;395(10225):689-697. [doi: 10.1016/S0140-6736(20)30260-9]

Abbreviations

API: application programming interface
China CDC: Chinese Center for Disease Control and Prevention
COVID-19: coronavirus disease

Edited by G Eysenbach; submitted 16.04.20; peer-reviewed by H Liang, KW Fu, E Lau, C Basch; comments to author 08.05.20; revised version received 18.05.20; accepted 25.05.20; published 28.05.20

Please cite as:
Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W
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 2020;22(5):e19421
URL: http://www.jmir.org/2020/5/e19421/
doi: 10.2196/19421
PMID: 32452804