Disaster Medicine and Public Health Preparedness

www.cambridge.org/dmp

Monitoring the Relationship between Social Network Status and Influenza Based on Social Media Data

Qi Yan PhD[1], Siqing Shan PhD[2,3], Baishang Zhang[4], Weize Sun MS[2,3], Menghan Sun MS[2,3], Yiting Luo MS[2,3], Feng Zhao PhD[2,3] and Xiaoshuang Guo MS[2,3]

[1]Management School, Tianjin Normal University, Tianjin, China; [2]School of Economics and Management, Beihang University, Beijing, China; [3]Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operation, Beijing, China and [4]Development Research Center of State Administration for Market Regulation of the PR China, Beijing, China

Original Research

Cite this article: Yan Q, Shan S, Zhang B, *et al*. Monitoring the relationship between social network status and influenza based on social media data. *Disaster Med Public Health Prep*. **17**(e490), 1–13. doi: https://doi.org/10.1017/dmp.2023.117.

Keywords:
social network status; physical condition; influenza; public health; information model

Corresponding author:
Siqing Shan; Email: shansiqing@buaa.edu.cn.

Abstract

Background: This article aims to analyze the relationship between user characteristics on social networks and influenza.

Methods: Three specific research questions are investigated: (1) we classify Weibo updates to recognize influenza-related information based on machine learning algorithms and propose a quantitative model for influenza susceptibility in social networks; (2) we adopt in-degree indicator from complex networks theory as social media status to verify its coefficient correlation with influenza susceptibility; (3) we also apply the LDA topic model to explore users' physical condition from Weibo to further calculate its coefficient correlation with influenza susceptibility. From the perspective of social networking status, we analyze and extract influenza-related information from social media, with many advantages including efficiency, low cost, and real time.

Results: We find a moderate negative correlation between the susceptibility of users to influenza and social network status, while there is a significant positive correlation between physical condition and susceptibility to influenza.

Conclusions: Our findings reveal the laws behind the phenomenon of online disease transmission, and providing important evidence for analyzing, predicting, and preventing disease transmission. Also, this study provides theoretical and methodological underpinnings for further exploration and measurement of more factors associated with infection control and public health from social networks.

Since the emergence of influenza (flu) began hundreds of years ago,[1] outbreaks of influenza virus have been periodic. In addition, because of its strong infectious qualities and its ability to mutate easily, a global influenza pandemic could result in severe human, economic, and social consequences. In the past decade, with the quicker growth of urbanization and population concentration, influenza pandemic has been the greatest threat to global public health.[2] These diseases can evolve and spread rapidly, greatly affecting people's health and happiness and even affecting the country's stability and security. In addition, these days, almost every time, the emergence of new or mutated virus has posed an enormous threat to people, as in the case of severe acute respiratory syndrome (SARS),[3] H1N1,[4] and coronavirus disease 2019 (COVID-19).[5] Meanwhile, these new situations have brought great challenges to traditional disease surveillance and infection control. Research on influenza is of great academic and practical significance. Thus, how we respond to these threats deserves further study.

Influenza has affected people's lives because of its great infectivity and variability. For that reason, flu control and prevention have received much attention from researchers and governments. Today, many traditional influenza prevention and control systems are in operation, such as the World Health Organization (WHO) Global Influenza Surveillance Network (FluNet), the US Outpatient Influenza-like Illness Surveillance Network (ILINet), and the European Influenza Surveillance Network (EISN). However, offline data are associated with problems such as a wide range of sources, long reporting cycles, and high costs. Data collection requires many resources to maintain, and it cannot be updated on a timely basis.

Our goal is to explore the association between the characteristics of the users on social networks and the susceptibility to influenza. In this study, we analyze and extract influenza-related information from all other information of users in social networks, which features many advantages such as efficiency, low cost, and real-time information. Additionally, we choose influenza-related Weibo updates of users to measure the susceptibility to influenza. Furthermore, we establish rankings on the basis of in-degree centrality to quantify the social network status of users. Then, we can explore the interplay between social network status and

© The Author(s), 2023. Published by Cambridge University Press on behalf of Society for Disaster Medicine and Public Health.

influenza. Also, we extract users' physical condition feature from social media based on LDA topic model to conduct its correlation analysis with flu. Throughout this process, different machine learning techniques are applied to classify influenza-related information separately from all other information in an automatic and effective manner.

The value of our research lies in revealing the laws behind the phenomenon of online disease transmission, and providing important evidence for analyzing, predicting, and preventing disease transmission. From the perspective of common medical knowledge, social network is considered irrelevant to disease transmission vectors because online interactions do not lead to people's direct contact. However, from the findings of our research, social media has become an invisible transmission vector for diseases in a sense. We speculate that the possible reasons include 2 points. (1) People with higher social media status have greater influence in social network interactions. Meanwhile, social media is also a vehicle for emotion transmission. A large number of studies proved that emotions may affect people's physical health.[6–8] (2) It means that people with high social media status have a large number of followers. Most often they are not only involved in frequent online activities, but are also invited to participate in corresponding social activities offline. It contributes to a high probability of being infected as well.

This research has important theoretical implications. It provides further evidence that there is a link between user characteristics in social networks and influenza. It also proves that Sina Weibo has great value in disease research. It is reliable to use social media data to conduct research related to epidemics and other health-related issues. Although there are several factors leading to distortion of social media data, we clean and filter the data during the experiment, and also compare different classifiers to exclude irrelevant data. Meanwhile, the use of machine learning further improves the accuracy and reliability of the data. The practical contributions appear in providing a classification model for detecting influenza-related Weibo messages. According to the results, users in social networks can be identified as belonging to different groups on the basis of their susceptibility to influenza. Therefore, we provide a scientific basis for public health intervention.

Throughout this article, from the perspective of social networking status, we divided the users into different groups to find vulnerable groups. In addition, we provided theoretical help to individuals who focus on maintaining their health. Therefore, we provided a basis for judgments for targeted health prevention and intervention measures and furnished better help to improve public health. In future work, we will dig more into the characteristics of users in social networks and explore the relationship between these characteristics and influenza. Our results will be useful for the allocation of influenza prevention resources.

Related Works

Due to the rapid advancement of Internet technology, the Internet has gradually become accepted as an important tool for modern people to access information and communicate. Meanwhile, as an important platform for people to use to express their opinions and emotions and interact with others, we can obtain user-generated content easily and in real time from social networks.

The emergence of social networks has provided an opportunity for health-related research.[9,10] Some studies attempt to make use of data from social media to support those systems[11] and also demonstrate the reliability of social media data.[12] Early prediction of seasonal epidemics like influenza can be enhanced by using social networking sites and Web blogs for real-time analysis, which enables faster tracking and better predictions compared with traditional methods. Social media data provide an efficient resource for disease surveillance and early warnings, offering an alternative solution to slow and expensive approaches like ILINet.[13] Scanfeld et al.(2010) analyzed Twitter status updates mentioning antibiotics to categorize their content and identify cases of misunderstanding or misuse. The results revealed various categories and instances of misunderstanding or abuse, particularly related to the combination of antibiotics with "flu" and "cold."[14] Aramaki et al. (2011) used the Twitter API to obtain flu-related tweets for correlation testing, which proved the feasibility of social media data to reflect the real world.[15] Twitter data have been used to predict the swine flu pandemic.[16] Yousefinaghani et al. (2019) collected and analyzed posts discussing avian influenza on Twitter to assess the Twitter's potential for outbreak detection, and the proposed approach was empirically evaluated using a real-world outbreak-reporting source. It is found that 75% of real-world outbreak notifications of AI were identifiable from Twitter.[17] Paul et al. (2014) showed that Twitter data can help reduce the forecasting error in ILI prediction and advance 2 to 4 wk ahead of baseline models.[18] Aiello et al. (2020) studied the public health tracking and prevention, addressing the importance of social media- and Internet-based data in disease surveillance.[19] Lampos et al. (2015) indicated that a nonlinear query modeling approach delivers the lowest cumulative nowcasting ILI rate error, and suggested that query information significantly improves autoregressive inferences, obtaining state-of-the-art performance.[20] The research of Masri et al. (2019) found that Zika tweets were a significant predictor of ZIKV cases, with model evaluation demonstrating that weekly ZIKV case counts could be predicted 1 week in advance. The results showed models using Twitter data are better predictors of Zika virus epidemic than models using traditional case report data.[21] Samaras et al. (2020) collected data on influenza in Greece from Google and Twitter and compared with influenza data from the official authority of Europe. The result shows that Google and Twitter both have potential to estimate and predict influenza but Twitter has some advantages over Google for that it achieves slightly better accuracy.[22]

Many studies have proven the correlation between social status and citizens' health. Hoebel et al. (2017) and Huynh and Chiang (2018) investigated the correlation between subjective social status and health (blood pressure, somatic symptoms, etc.) in adults and adolescents, respectively.[23,24] Stringhini et al. (2017) introduced 25×25 risk factors into the model to observe the contribution of socioeconomic status and these conventional risk factors to mortality and life loss.[25] Not only physical health, but also social status has an impact on citizens' mental health.[26,27] Euteneuer et al. (2021) used panel data to provide new insights into the longitudinal pathway of subjective social status and health.[28] In addition to subjective social status, objective socioeconomic status has also been shown to correlate with physical health.[29] Moreover, unequal socioeconomic status in different regions and communities has different effects on physical and mental illness.[30–32] According to research, immune function and social status in the real world were found to be directly correlated.[33]

In addition, some research shows that social network data are also used to study social relations and activities of users in that social network. For example, data from social media can be used to reveal influenza transmission based on the users' location and

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

social ties.[34] Perez-Rodriguez et al. (2020) used data from social media to build models, revealing the impact of social status, exposure to pollution, and many lifestyle factors on one's health.[35] This research shows that the characteristics or behaviors of a user within a social network are likely to be relevant to their health, and different individuals or groups may have different health conditions. Murayama et al. (2021) proposed a method using social media and commuting data to predict the geographical distribution of influenza patients and validated the accuracy of the predictions against weekly influenza patient data from health authorities, serving as ground truth.[36] Qin and Ronchieri (2022) analyzed a large dataset of tweets related to various pandemics and explored natural language processing techniques to extract insights from unstructured text comments, revealing that discussions primarily focused on malaria, influenza, and tuberculosis, with prevalent emotions of fear, trust (specifically related to HIV/AIDS), and disgust.[37] Wang et al. (2022) investigated the mobile social media dissemination behavior of the public and found that national-oriented risk culture and strict scrutiny of social media influenced mobile social media users' seeking and sharing of disease information during public health emergencies.[38] However, most of the studies focus on the prediction of flu trends. There are relatively few influenza-related studies targeted at the differences among different groups in social networks. Social rank can influence the health of an individual, particularly with respect to stress-related disease, but the relationship between social status and infectious disease still needs more study and verification.[39] The research of Okamoto et al. (2011) supports a negative association between social network status, in-degree centrality, and depressive symptoms.[40] Further comparison with the latter 2 parts of the literature is presented in Table 1.

Methods

This section describes the methods we use to automatically identify influenza-related information (indicating that the user has caught influenza) authored by users and to quantify the susceptibility of people to influenza, the social network status, and the physical condition of users. First, we screen influenza-related data on the basis of Weibo messages. Then, for each selected unique user, we mine the total number of Weibo messages, the exact count of influenza-related Weibo, and the in-degree of nodes to obtain the measurement of the index. Finally, we explore the relationship between the social network status of users and their susceptibility to influenza and also study the relationship between the physical condition and susceptibility to influenza. We will walk through these steps in detail in the upcoming subsections.

Modeling the Detection of Influenza-Related Information

We built a model to apply to Chinese short text classification such as the contents of Weibo. We used 6 different classifiers to find a method that had the best performance. Machine learning consists of supervised (classification and regression) and unsupervised (clustering and generalization) learning but also semi-supervised and ensemble learning.[41] The classifiers used in this study include k-nearest neighbor, decision tree, SVM, naïve Bayes (NB; multinomial model and Bernoulli model). These classifiers help us to distinguish between Weibo messages indicating that the user is suffering from influenza (label 1) and all other information (label 0).

Before the classification, we outlined some steps to process these data. First, in terms of preprocessing, this study contains 2 approaches: the first 1 is Chinese word segmentation, and the second is removing stop words. The stop words used in this study are sourced from the stop words list released by the Natural Language Processing Laboratory at Harbin Institute of Technology. Second, we use information gain (IG) to extract the features that benefit classification. IG is an important index in the selection of features. A feature is important for bringing more information to the classification system. The most popular feature selection method is IG, which works well with texts and has often been used.[42] Finally, according to the features, we complete the vectorization of texts by using TF-IDF.[43] The TF-IDF algorithm is always used to weigh each word in the text according to how important it is, and it captures the relevancy among words, text documents, and particularities.[44]

Then, we use machine learning techniques to distinguish texts and filter useless texts. The entire process is shown in Figure 1.

Modeling Correlation: Social Network Characteristics and Susceptibility to Influenza

In this section, we quantify the social network characteristics of users, including social network status and physical condition, as well as the susceptibility to influenza. We quantify the individuals' susceptibility to influenza using a ratio of the quantity of the user's influenza-related Weibo messages divided by the user's total number of Weibo messages. In a social network, users will share their health-related information, which can be used to infer health status and incidence rates for specific conditions or symptoms.[45,46] First, through the classifier in the previous step, we obtain the Weibo messages that describe authors with influenza. Then, we mine all Weibo messages of these authors containing the keywords "influenza" (gan mao).

Due to the differences in social media platforms and language characteristics, the processing of Chinese text differs from other languages. First, Chinese words are typically represented at the character level, unlike English words separated by spaces. Therefore, in Chinese text processing, it is necessary to perform word segmentation, dividing continuous sequences of characters into meaningful words. Word segmentation is a critical step in Chinese text processing and plays an important role in subsequent text analysis and semantic computation. Second, in the context of Chinese expression, descriptions of influenza symptoms by patients may have certain colloquial features. This is because colloquial expressions are closer to everyday conversations and real-life language usage situations. Moreover, on social media platforms, users tend to use colloquial language styles to express their emotions and feelings. The degree of colloquial features used in Chinese Weibo texts may vary depending on individual differences and the characteristics of social media platforms. Therefore, when analyzing Chinese Weibo texts, this study combines feature extraction techniques and language models to capture and understand these colloquial expressions, accurately extracting and analyzing information related to influenza symptoms. Table 2 shows the raw crawled data information aggregated in user dimensions.

We use these data as inputs for the classifier to screen out valuable influenza-related data. We also need to count the number of influenza Weibo messages of each person. In addition, we obtain each individual's total number of Weibo messages. Regarding social network status, this article differentiates these users by

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

Table 1. Comparison of previous research and proposed work

Authors	Objective	Study outcome	Specific contributions of this work
Hoebel et al. (2017)	Research on relations between social status and health.	Perceptions of low social status may be related to the pathogenesis of depression among adults in Germany.	These research findings consistently point toward the relationship between subjective social status or socioeconomic status and health. In addition, few studies have focused on physical illness, specifically the flu. Our study investigates how individuals perceive their social network influence and objectively quantify their social media status, rather than social status in reality. In addition, we verify the coefficient correlation between social media status and health condition.
Huynh et al. (2018)		Subjective social status may affect adolescent health.	
Stringhini et al. (2017)		People with low socioeconomic status have greater mortality compared with those with high socioeconomic status.	
Fournier (2020); Uecker and Wilkinson (2020)		Not only physical health, but also social status has an impact on citizens' mental health.	
Euteneuer et al. (2021)		Both local and national subjective social status have associations with health.	
McMaughan et al. (2020)		Objective socioeconomic status is related to health-care access as well as health-care access and health outcomes for older adults.	
Kivimäki et al. (2020)		Low socioeconomic status is associated with increased risk for 18 (32·1%) of the 56 common diseases and health conditions.	
Wanberg et al. (2020)		People with higher education experience a greater increase in depressive symptoms from before to during COVID-19. The relation between socioeconomic status and child psychopathology may vary in different populations of children and in different communities.	
Peverill et al. (2021)		Revealing the impact of social status, exposure to pollution and many lifestyle factors on one's health based on social media data	
Tung et al. (2012)		Social status is associated with gene expression levels.	
Hassan et al. (2019)	Research on social relations and activities of users in social network.	Location-based social media can be leveraged to monitor flu outbreaks.	These studies use social network information such as geolocation information and social interaction information to detect flu outbreaks or analyze its relationship with stress-related diseases. This study examines not only social network status but also physical condition from social media dataset to reflect their relationships with influenza.
Perez-Rodriguez et al. (2020)		Social interactions on Twitter can help identify and analyze the behaviors, perceptions and appreciations of patients and close relatives toward a health condition.	
Murayama et al. (2021)		Using social media and commuting data to predict the geographical distribution of influenza patients	
Qin et al. (2022)		Revealing that discussions primarily focused on malaria, influenza, and tuberculosis, with prevalent emotions of fear, trust and disgust	
Wang et al. (2022)		Investigated the mobile social media dissemination behavior of the public and found that national-oriented risk culture and strict scrutiny of social media influenced mobile social media users' seeking and sharing of disease information	
Sapolsky (2004)		Social rank is related to stress-related disease.	
Okamoto et al. (2011)		Social network centrality is significantly associated with depression among adolescents.	

relying on the in-degree centrality. Centrality is a concept commonly used in social network analysis (SNA) and is an attribute of nodes (users in the network) that is used to quantify network locations of nodes.[47] Centrality is considered to be a structural attribute of social networks in this study and is, therefore, widely used as an indicator to measure the importance of nodes in the network.[48–50] Degree Centrality was formally proposed for the first time in the paper by Linton C. Freeman.[51] It is one of the basic measures, suggesting the sum of 1 node directly connected to other nodes and is divided into in-degree centrality and out-degree centrality when the connection is directional. In-degree centrality represents other nodes connecting to a particular node.[52] It always relates to studies of popularity.[53] For the characteristics of social relations in Weibo, as a large and complex network, it is difficult to obtain centrality data. If building a small network, the results may be inaccurate due to the small sample size. Therefore, we consider the in-out degree concept in the complex network to measure the social network status. We regard individuals with more followers as people who have high social network status, while people who have low social network

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

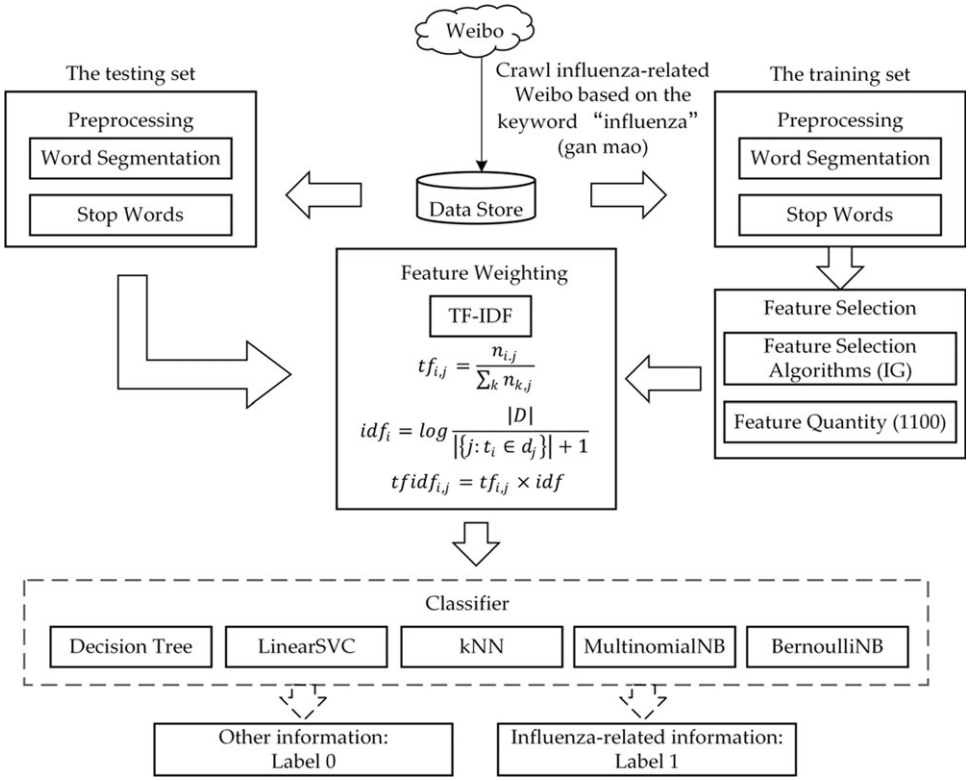


Figure 1. Model of detection of influenza-related information.

status are defined as individuals with fewer followers. As is shown in Table 3, there is a high standard deviation in followers, adversely affects subsequent correlation analysis. According to Likert scale, 5 ordered levels are often considered as the level of variable scales measurement,[54] so we determine 5 social ranks to describe the position of users in Sina Weibo.

From Table 4, it can be observed that although gender and region do not have a significant impact on the frequency of updates, there is a notable difference in sample size between males and females. The sample size of females is 4 times larger than that of males, indicating that females are more likely to post influenza-related information when experiencing flu symptoms. Additionally, there is no significant difference in data volume between northern and southern regions, suggesting that regional factors may not play a prominent role in the frequency of updates related to influenza.

Table 5 shows the distribution of users' followers with different in-degree levels. The social ranks from 1 to 5 indicate the increase in in-degree. The user's in-degree reflects the user's social influence. Users' in-degree is approximately long-tail distribution, which indicates that influential users in a social network are always in the minority.[55] If the number of followers is less than 100, the social rank is 1, and if the number of followers is between 100 and 1000, the social rank is 2. Table 5 provides details of the definition.

After obtaining these indicators, we study the relationship between social network status and susceptibility to influenza using the rank correlation coefficient (Spearman). We use the Spearman correlation coefficient because it has less strict requirements on the data conditions as long as the observations of the 2 variables are paired rank ratings data, or rank data converted from continuous variable observation data.[56,57] The rank correlation coefficient is given by

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (1)$$

where ρ_s is the Spearman rank relational coefficient, d_i is the difference between x_i' and y_i', x_i' and y_i' represent the position of the original data in a sorted sequence.

In addition, we also quantify the physical condition of users to measure its relationship with influenza susceptibility. First, before modeling, all the Weibo texts of target users experience preprocessing, word segmentation, and de-stopping words, etc. Second, the topic distribution and keywords are obtained by LDA topic model. Then we set 50 topics to output for LDA model, as each topic contains 10 keywords with probability. Finally, the topic related to physical condition are selected and summarized from the results of LDA. Meanwhile, 4 representative keywords under the topic are screened out and listed in Table 6.

Next, the Word2vec[58] model is adopted to carry out word clustering, finding the similar words of the existing 4 representative keywords under each topic. We choose the CBOW model and a 300-word vector dimension. In addition, words with frequency less than 2 are ignored. The training data of Word2vec is consistent with the corpus used in LDA training. After training, the top 10 words with the highest similarity of each given keyword are reserved. We remove the duplicated words from all of the similar words to obtain the final dictionary. The results of clustering similar words by Word2vec are shown in Table 6.

After expanding the keyword library, we match all the keywords in Weibo texts. Once the keywords of the corresponding topic appear in a text, it will be marked. Finally, we count the number of Weibo texts with corresponding keywords as the topic score of each target user and calculate the ratio of the topic score and the total number of Weibo texts to represent the variable of the topic

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

Table 2. Raw crawled data field

Categories	Crawled data field
Users' information	Blogger name, blogger ID, number of Weibos, gender, age, place of registration
Social network relationship	No. of followers, number of followings, first 5 page following list (homepage, blogger name)
Weibo texts	All users' Weibo texts entries and all users' Weibo texts containing flu keywords in 2017

Table 3. Descriptive statistics of followers

Variable	Min	Max	Mean	Standard deviation	Variance
Followers	1	1393991	4161.28	58064.641	3.37E+09

Table 4. Descriptive statistics of gender and region of user susceptibility

Categories	Number	Mean	Standard deviation	Standard error
Female	964	0.00723	0.0349	0.0011
Male	341	0.00986	0.0238	0.0016
Northern Region	745	0.00703	0.1647	0.0007
Southern Region	560	0.00715	0.1406	0.0007

Table 5. Division of social network status

In-degree (no. of followers)	Rank
1~100	1
100~1000	2
1000~10000	3
10000~100000	4
More than 100000	5

Table 6. Results of word clustering by Word2vec

Physical condition keywords	Similar words
Sick	crying, diarrhea, low fever, cough, catch a cold, get better, appetite, dementia, spleen and stomach, invigorate the brain, healed, orexis, intelligence, disease, runny nose, inflamed, sore throat, headache, rhinitis, coryza, cool, sultry, too hot, cooling, dryness-heat, damp, gloomy, burning hot, cold hands and feet, turn cold, resistance, feel hot, weather, get cooler, get cold, cough up, liver pain, spasms, nosebleed, neurasthenia, toothache, upset stomach
Chemotherapy	
Immunity	
Fever	

preference. Because the variable under the topic of physical condition is continuous, we compute its Pearson correlation coefficient with the susceptibility to influenza. The Pearson correlation coefficient is given by,

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (2)$$

where $r_{X,Y}$ is the Pearson correlation coefficient, $cov(X, Y)$ is the covariance of variables X and Y, σ_X and σ_Y represent the standard deviation of X and Y, respectively.

In the following sections, the experiment and results on the basis of our models will be introduced in detail.

Results

This section describes the experiments. We follow the model set up in the previous sections to obtain the results.

This work is based on data gained from Sina Weibo, one of the most popular social media platforms in China. According to Sina's first-quarter results of 2017, Sina Weibo now has more than 340 million active users worldwide and has surpassed Twitter. Sina Weibo is China's largest microblogging service.[59] Weibo allows users to post 140-character messages. Similar to Twitter, relationships between users on Twitter are not necessarily symmetric. Users can follow friends or interesting users without being followed back.

Using crawler software, we collect a sample of Weibo messages based on the keyword "influenza" (gan mao). We follow the principle of privacy protection, and all information obtained from crawler software is public information. We generate a random number behind each piece of Weibo data, rearrange it using random number sorting, and select 100+ pieces of data as the result of random sampling every month from January to December 2017. Finally, a total of 1305 unique users are used as samples for analysis (see Table 7). Next, we crawl all the public posts of the 1305 users, a total of 550,000 pieces of data generally. Because this work studies individuals' susceptibility to influenza and is about the vulnerability of the users, we also have a second mining of all of the tagged user's influenza-related Weibo messages.

In the text classification stage, 1760 records selected randomly from all crawled Weibo messages are used as training samples and testing samples. The dataset was assigned to 8 humans to label. Before officially labeling, we conducted prelabeling training for the 8 people and provided them unified labeling rules. In the process of officially labeling, each member was required to label 220 records according to unified labeling rules without communication to intentionally make the marked category accurate. After labeling, these data are divided into influenza-related information and all other information. Table 8 shows the tagged Weibo text. In this table, we show 6 messages from 6 unique users on Sina Weibo. Label 1 represents that these texts accurately describe a user who suffers influenza. Conversely, Label 0 indicates that the texts do not reflect the user's health. Next, we categorize 1320 records as the training set and the remaining records as the testing set. To compare and select the most appropriate classifier, we use 4 different kinds of classifiers, including NB, SVM, decision tree, and kNN. In this process, we also use IG to select a different number of features to determine the best combination of feature dimensions and classifiers. The indicators we applied to evaluate the experimental result of each classifier are accuracy, precision, recall, and F1-score. These 4 indicators are commonly used evaluation indicators for machine learning classification algorithms.[60] Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined, which means it is an important statistical measure of

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

Table 7. Summary statistics of the data and collected field

Category	Content
Users total	1305
Number of Weibos	550,000
Keyword	influenza (gan mao)
Field	Text, followers_count, Weibo_count, created time

Table 8. Tagged Weibo text

Text	Label
The strong wind these days finally have made me catch influenza . . .	1
Influenza made me dizzy and top-heavy. I did not feel like eating anything.	1
I have got influenza and a bad cough for days. Does anybody have a good idea to make me feel better?	1
People's feelings are like influenza. No matter how hard you are trying to get over it, it takes a few days to cure.	0
From the patients' perspective, it costs them approximately 200-300 yuan to see a doctor for influenza. How dare you say it is not expensive at all?	0
It takes only 7 d to get over influenza. However, you are the only person I have been thinking about in the past 7 y.	0

how well a binary classification test correctly identifies an item.[61] The precision, recall, and F1-score, which are usually used together, are also other important indicators to measure classifier performance. The F1-score is a comprehensive consideration of precision and recall.

Figure 2 illustrates the results of detecting influenza-related information. The horizontal axis is the number of features (100-1100). The experiment started with selecting 100 features and ended with 1100 features. The vertical axis represents the accuracy, precision, recall, and F1-score, respectively. The machine learning methods used in this model are described in the rectangle in the upper left corner. Each line in Figure 2 represents a machine learning method.

As shown in Figure 2(1), the accuracy of kNN and decision tree performance are both lower than 0.6. The performance of the 2 modes of NB is relatively good, but when the number of features is less than 600, the difference becomes obvious, and the Bernoulli NB performs stably. In terms of precision in Figure 2(2), we can see that the multinomial NB maintains the highest precision in the process. The Bernoulli NB performs better when the number of features increases, while the kNN and decision tree still underperform. With the increase in features, the accuracy and precision of kNN decline. Other classifiers meanwhile increase their performance as the number of features increases. From the perspective of the recall shown in Figure 2(3), the performance of NBs is still the most stable, and it is worth noting that when the number of features is less than 600, the performance of SVM (linear SVC) is better than that of multinomial NB, whose recall reaches 0.6-0.7. From the F1-score results in Figure 2(4), Bernoulli NB maintains the best performance.

In summary, Bernoulli NB shows the best performance in the process, suggesting the NB has obvious advantages in this binary classification and a stable performance. Therefore, we choose 1100 features and use Bernoulli NB for classification. The model has the best performance. The accuracy reaches 80.68% and the F1-score reaches 80.66% on the testing set. The complete classifier performance statistics are shown in Table 9. To obtain the optimal result, the IG is used for feature selection, with a dimension of 1100, and Bernoulli NB is applied to classify all the Weibo messages.

We also compare the best-performing machine learning model (ie, Bernoulli NB) in our study with several deep learning models from reference,[62–64] and the comparative results are presented in Table 10. It can be observed that, due to the limitation of data availability, the performance of the deep learning models did not surpass that of the Bernoulli NB.

According to the above results, we use Bernoulli NB to classify all text, and we obtain the total number of influenza-related Weibo messages from each individual by using this classifier. As stated earlier in this article, we can express the individuals' susceptibility to influenza by using this ratio. Regarding social network status, according to the method in the model, we divide these data into 5 social ranks. Then, we randomly select 100 records from every social rank for correlation analysis. Table 11 shows descriptive statistics of individuals' susceptibility to influenza in each rank (1-5).

Each rank contains 100 users. Rank 1 means the in-degree centrality of users between 1 and 100 and users of rank 5 have more than 100,000 followers. As we can see in Table 11, the mean of influenza susceptibility decreases with increasing rank. Although there are no obvious trends at ranks 2 and 3, this situation may be caused by extreme values determined from the maximum, minimum and standard deviation. However, it does not affect the overall trend.

Figure 3 shows the association between the measure of individuals' susceptibility to influenza and social network status. From the chart, we can see that groups that possess different numbers of followers in the social network have different results of the susceptibility to influenza. With the increasing quantity of followers (horizontal axes), individuals' susceptibility to influenza decreases (vertical axes). The outcome is consistent with the results shown in Table 11. For example, in the group of individuals with 1-100 followers, we can infer that they are very vulnerable to infection. In addition, in the group of individuals with more than 100,000 followers, we can easily discover that they are less susceptible to influenza because these points are very concentrated and close to zero. Applying the rank correlation coefficient, we quantified the relationship between individuals' susceptibility to influenza and social network status, and we found a moderate correlation and a significant negative correlation, with a correlation coefficient of -0.427 ($P < 0.001$).

As for the physical condition topic identified from the Weibo texts, we carry out Pearson correlation coefficient to analyze. As shown in Table 12, when users focus on their physical condition, their susceptibility to influenza and corresponding topics show a significant positive correlation. It suggests that users who post more Weibo texts about physical condition may have a higher flu susceptibility. Through the observation of the keywords in the topic, the Weibo texts describing the physical condition are mostly related symptoms and sensitive emotions to the physical condition.

In general, regarding the 3 research questions addressed in this study, the following conclusions can be drawn. First, in identifying and classifying influenza-related tweets, the Bernoulli NB model achieved better classification performance with an accuracy of 0.8068. The proportion of a user's influenza-related tweets can be used as an indicator to calculate influenza susceptibility. Second, the experimental results indicate a moderate negative correlation between social status and influenza susceptibility. This means that individuals with higher social centrality have a lower susceptibility to influenza. Third, there is a significant positive correlation

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

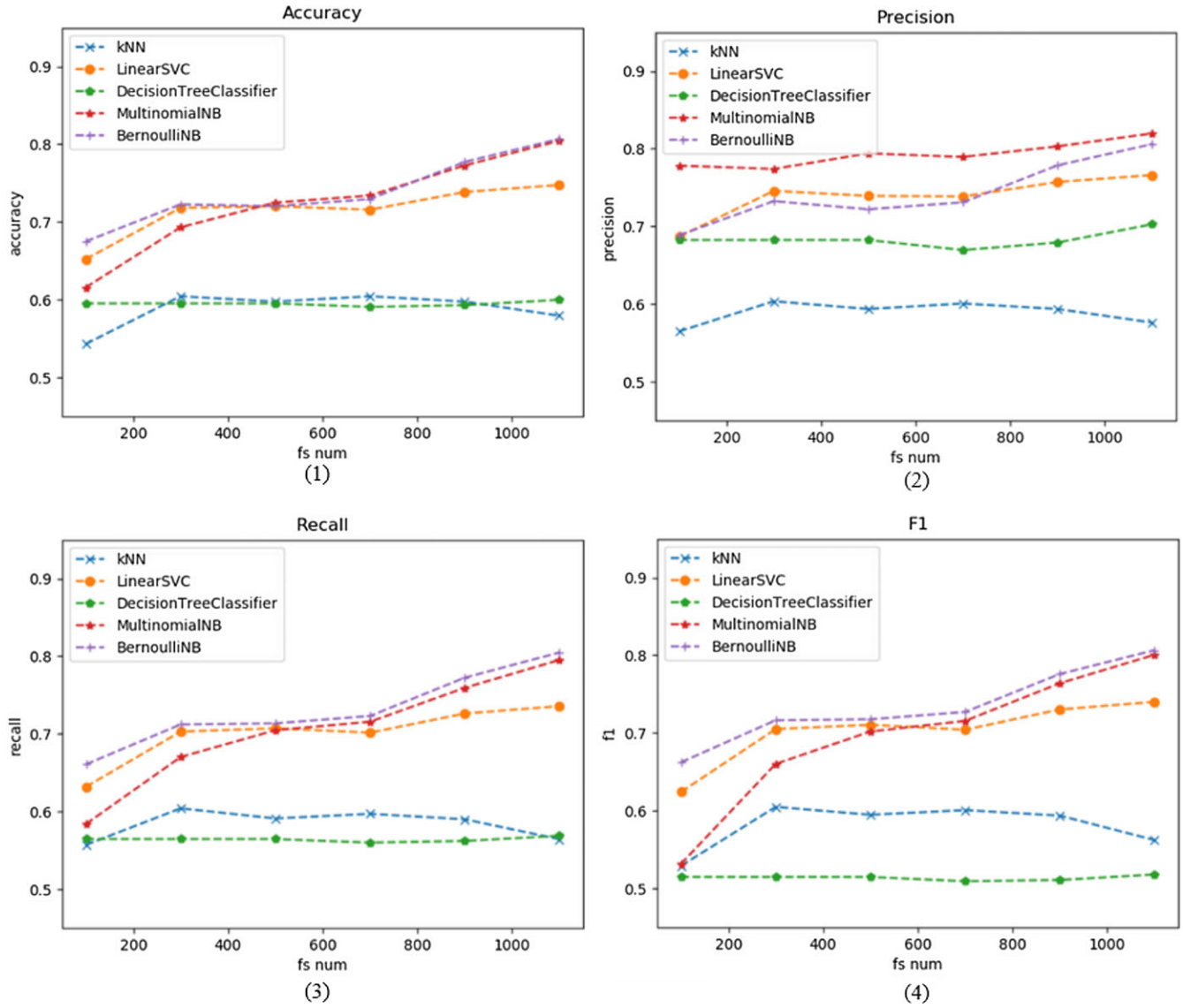


Figure 2. Performance of different classifiers.

between the user's reported physical condition on social media and influenza susceptibility. These findings contribute to a better understanding of the relationship between social media, individual characteristics, and influenza susceptibility. The results highlight the potential of social media data in studying and predicting the spread of influenza and provide insights for public health interventions and prevention strategies.

Discussion

In this study, we choose Sina Weibo for our research. Sina Weibo is one of the most influential and popular social media platforms in China. According to the latest first quarter of 2017 results for Sina Weibo, as of March 31, the number of Sina Weibo monthly active users reached 340 million, meaning the platform has overtaken Twitter as the world's largest independent social media company. In China, as one of the most popular social media platforms, Sina Weibo is also used for scientific research.[65] For example, Xu et al. (2019) trained a classifier to identify and detect rumors from a mixed set of true information and false information.[59] Chen et al. (2020) found that sentiment influences the retweet patterns and retweet speed of social media.[66] However, disease-related research is relatively rare, and most of these studies are targeted for surveillance of infectious disease.[67–69] In addition, there is research revealing significant differences in the microblogging behavior on Sina Weibo and Twitter.[70] One significant difference lies in the language patterns used on Weibo and Twitter. While English is predominantly used on Twitter, Weibo is a Chinese microblogging platform where the majority of content is in Chinese. This language distinction has implications for text analysis and natural language processing techniques applied to social media data. For instance, the Chinese language relies on character-based representation, necessitating the use of segmentation techniques to identify meaningful words or phrases within a continuous stream of characters. User behavior also differs between Weibo and Twitter. Weibo users tend to engage in more active and frequent interactions, often using various multimedia formats such as images, videos, and emojis to convey their messages. On the other hand, Twitter users may focus more on concise and succinct expressions due to the platform's character limit. These differences in user behavior influence the content and

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

Table 9. Performance statistics comparison of different classifiers

Classifier	Feature dimension	Accuracy	Precision	Recall	F1-score
kNN	100	0.5431	0.5751	0.5572	0.5292
	300	0.6045	0.6037	0.6043	0.6051
	500	0.5977	0.5938	0.5912	0.5950
	700	0.6045	0.6009	0.5972	0.6009
	900	0.5977	0.5937	0.5902	0.5939
	1100	0.5795	0.5763	0.5645	0.5624
SVM (linearSVC)	100	0.6522	0.6868	0.6320	0.6242
	300	0.7181	0.7458	0.7031	0.7052
	500	0.7205	0.7393	0.7072	0.7107
	700	0.7159	0.7386	0.7017	0.7043
	900	0.7386	0.7573	0.7263	0.7304
	1100	0.7477	0.7661	0.7358	0.7402
Decision tree	100	0.5955	0.6825	0.5648	0.5149
	300	0.5932	0.6730	0.5626	0.5133
	500	0.5932	0.6730	0.5626	0.5133
	700	0.5932	0.6793	0.5623	0.5111
	900	0.5955	0.6894	0.5644	0.5127
	1100	0.6000	0.7029	0.5690	0.5182
Multinomial NB	100	0.6159	0.7782	0.5841	0.5318
	300	0.6932	0.7740	0.6703	0.6604
	500	0.7250	0.7940	0.7048	0.7020
	700	0.7341	0.7894	0.7157	0.7157
	900	0.7727	0.8030	0.7593	0.7641
	1100	0.8045	0.8197	0.7952	0.8007
Bernoulli NB	100	0.6750	0.6887	0.6609	0.6624
	300	0.7227	0.7325	0.7122	0.7165
	500	0.7205	0.7221	0.7137	0.7178
	700	0.7295	0.7311	0.7231	0.7272
	900	0.7773	0.7787	0.7724	0.7760
	1100	0.8068	0.8060	0.8048	0.8066

Table 10. Performance statistics comparison with deep learning classifiers

Classifier	Accuracy	Precision	Recall	F1-score
TCNN-Bi-LSTM[62]	0.6667	0.6875	0.8614	0.7647
Extra tree with TF-IDF and BOW feature combination[63]	0.6705	0.6872	0.8735	0.7692
ARCNN[64]	0.6629	0.7584	0.6807	0.7175
Bernoulli NB with 1100 feature dimension	0.8068	0.806	0.8048	0.8066

style of conversations, as well as the types of information shared on each platform. Furthermore, cultural influences play a significant role in shaping the dynamics of social media in Chinese contexts. Chinese culture values collectivism and group harmony, which can be reflected in the way users communicate and interact on Weibo. Users may emphasize consensus-building, social connections, and shared experiences, which in turn affect the topics discussed and the sentiment expressed on the platform. Understanding these cultural influences is crucial for interpreting social media data accurately and comprehensively. Therefore, Weibo data merit in-depth analysis to determine the link between disease and social network status.

Table 11. Descriptive statistics of susceptibility to influenza

Rank	N	Minimum	Maximum	Mean	Standard deviation
1	100	.000026	.333333	.022889	.039988
2	100	.000215	.076923	.006471	.009028
3	100	.000116	.111111	.006917	.014166
4	100	.000009	.031915	.004083	.005970
5	100	.000013	.014851	.002333	.003090

Although previous studies have shown that people's social status can affect their health, there is not much evidence of the relationship between social status in networks and influenza susceptibility, which needs to be further explored. In this study, we distinguish between individuals in the social network from the aspect of social network status and study the differences in susceptibility to influenza among different groups. Before this step, we obtain classification results by developing a Bernoulli NB classifier. This is a necessary precondition for further progress, as false labels will disturb the experimental results. Next, we showed a moderate negative correlation ($R = -0.427$; $P < 0.001$) between susceptibility to influenza and social network status through rank correlation analysis. This result indicates that people with higher social network status are more likely to have lower susceptibility to influenza. From this result, we can see that people who have low in-degree centrality in Sina Weibo (i.e, people with low social network status) may be more susceptible to influenza, although we cannot accurately explain this tendency through medicine. However, it has been proven in the previous literature that people with lower social rank are more prone to stress-induced illness. In addition, in real life, people with lower social status have less time and energy to take measures to care about their health. Through this experiment, we have also demonstrated that, to a certain degree, social status in the network also reflects social rank in real life. Furthermore, we conduct another experiment on the correlation between users' physical condition and their susceptibility to influenza. Pearson correlation coefficient showed that there was a significant positive correlation ($R = 0.915$; $P < 0.001$) between them. It means that people who prefer to share their physical condition on social media tend to receive a higher susceptibility. In real life, human immunity and the disorder of metabolism are typical of people's physical condition, and such characteristics are also reflected in the health-related information of social media.

This study explores the relationship between different network characteristics and influenza susceptibility by identifying infectious users in social networks and distinguishing them from different feature dimensions. Through this research, we can identify people who are susceptible to influenza and consider them key populations for early monitoring and control to avoid further spread of the epidemic. The experimental results show that individuals with low in-degree centrality are key targets for influenza monitoring, reminding and encouraging high-susceptibility populations to pay attention to their own health. Moreover, users who are already showing symptoms of diseases also have higher susceptibility to influenza. At the same time, this study explores a reasonable relationship between social network characteristics and influenza. This is an important step in identifying health-related factors from social networks, and it also provides a theoretical basis for public health surveillance. In addition, we find that social media data, as a "social sensor", are a supplementary information source for disease-related research and have crucial research value.

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

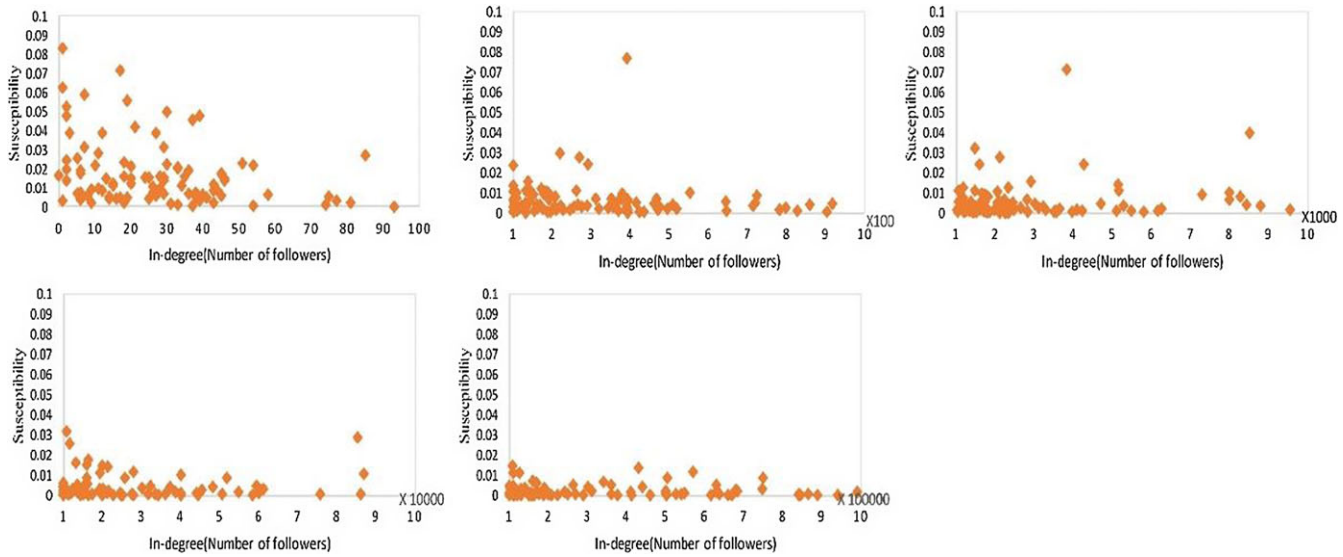


Figure 3. Association between susceptibility and social network status.

The innovation of this study, compared with previous research, is mainly reflected in the following aspects. First, this study focuses on the relationship between individuals' social media status and influenza susceptibility. Previous studies may have overlooked individual-level differences and susceptibility factors. This study aims to investigate the influence of individual-level differences in social networks on influenza susceptibility. This individual-focused approach provides us with a deeper understanding and helps uncover individual contributions to influenza transmission within social networks. Second, this study uses machine learning algorithms for text classification to ensure accurate identification and selection of Weibo posts describing users' own illness conditions from randomly crawled Weibo texts. By avoiding "zombie users," ie, inactive users, we can ensure the authenticity of the data and the users' activeness. Third, this study quantifies influenza susceptibility using social media data. Unlike previous studies, we not only consider the number of influenza-related Weibo posts in quantifying susceptibility but also introduce ratios as a measure. The advantage of this approach is that it eliminates the influence of users' posting frequency and habits on social networks. By comparing the number of influenza-related Weibo posts with the total number of posts made by users on social media, we can more accurately quantify individual influenza susceptibility.

The significance of the 3 research questions can be summarized as follows. The significance of Question 1 lies in the accurate identification of influenza-related information through machine learning algorithms applied to Weibo texts. This provides a reliable data foundation for further research on influenza transmission. By proposing a quantitative model for influenza susceptibility, it becomes possible to quantify users' overall susceptibility to influenza and explore factors related to influenza transmission at the social network level. The contribution of this research question lies in providing an effective method for identifying influenza-related information and establishing a quantitative model for influenza susceptibility in social networks. The importance of Question 2 lies in evaluating users' status and influence in social networks by using the in-degree metric from complex network theory and analyzing its correlation with influenza susceptibility. This helps us understand the impact of status and influence in social networks on influenza transmission. By validating the correlation coefficient between the in-degree metric and influenza susceptibility, further confirmation of the association between social network status and influenza can be obtained, providing new insights into the mechanisms of influenza transmission. The significance of Question 3 lies in the application of the LDA topic model to extract users' physical condition information from Weibo texts and analyze its correlation with influenza susceptibility, which reveals the potential impact of health conditions on influenza transmission. This approach can help us explore users' health conditions from a social media perspective, providing important reference for influenza prediction and intervention measures.

The potential value and applications of this research are multifaceted and impactful. First, studying the association between social network status and influenza susceptibility can help predict the spread of influenza within social networks. Key individuals or highly connected individuals within social networks may play crucial roles in the transmission of influenza. Understanding the relationship between social network status and influenza susceptibility can assist in identifying high-risk groups and targeted interventions, enabling better prediction and control of influenza transmission. Second, the findings from the research on social network status and influenza susceptibility can be used for health education and promotional activities. By understanding the characteristics and status of individuals with higher susceptibility within social networks, tailored health education measures and communication strategies can be developed. This can enhance awareness of influenza, strengthen preventive measures, and encourage individuals to adopt appropriate prevention and protection measures. Last, studying the relationship between social network status and influenza susceptibility can provide guidance for social media monitoring and information dissemination. Understanding the impact of social network status on the spread of influenza-related information can help design more effective information dissemination strategies, targeting individuals with different social statuses for directed communication and interventions. Additionally, leveraging social media platforms to

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press

Table 12. Pearson correlation coefficient between physical condition and influenza susceptibility

Variables	Correlation coefficient	Significance (2-tailed)	Result
Physical condition Influenza susceptibility	0.915	0.000(<0.001)	Significant positive correlation

provide real-time influenza information can facilitate better monitoring and response to influenza outbreaks.

In summary, studying the association between social network status and influenza susceptibility has implications for predicting and controlling influenza transmission within social networks, guiding health education and promotional activities, and improving social media monitoring and information dissemination strategies for influenza.

Conclusions

The H1N1 influenza pandemic occurred in 2009. By the end of 2009, at least 12,000 people had died because of the H1N1 influenza virus. Although we have a well-established monitoring system and vaccines to prevent and control the spread of this virus, the acceleration of population growth and urbanization has also introduced new factors in controlling the spread of influenza. The increase in population density and structural changes have greatly increased the probability of the transmission of infectious diseases in cities. Meanwhile, traditional influenza research is mainly based on traditional monitoring data. Most of the data come from hospitals and laboratories and are associated with high costs. Moreover, there is a lag in the collection of information. With the rapid development of social networks, people's lifestyles and information sources have undergone great changes. Today's social networks often have hundreds of millions of users interacting with each other and expressing themselves. Social networks are a new data source, and this development presents an important opportunity for influenza-related research and infection control.

Regarding the theoretical contributions, this article focuses on mining data from social networks. In recent years, access to public information on social networks has become very popular. There is also much research that proves that social network data are a reliable source of information, but most of it is foreign-related research. This research proves that Sina Weibo, as one of most popular social networks in China, also has great value in disease-related research. This experiment demonstrates a reasonable connection between social network status and influenza susceptibility, which is an important step in mining health-related factors from social networks. On this basis, we can put forward future studies in the field of disease. Our research is not just theoretical; it also has practical implications. In this study, we performed some disease-related work. Traditional health-related research is time-consuming and requires much human effort to collect data. Therefore, traditional methods are not conducive to timely disease surveillance and control. However, through social networks, we can easily and efficiently access users' information and without requiring the active participation of individuals. Therefore, we cannot only improve the reliability of the collected data but also lower costs by reducing the intermediate steps. Through the results of this study, individuals of low social network status can be identified as the key targets of influenza surveillance, and we also provide a theoretical basis for the public. Moreover, these results contribute to encouraging individuals with high influenza susceptibility to pay attention to their health.

Data availability. The data presented in this study are available on request from the corresponding author.

Acknowledgments. The authors gratefully acknowledge the support of Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. In addition, the authors thank the anonymous reviewers for insightful comments that helped us improve the quality of the study.

Author contributions. Conceptualization, Qi Yan and Siqing Shan; Data curation, Siqing Shan and Yiting Luo; Formal analysis, Qi Yan and Menghan Sun; Funding acquisition, Siqing Shan; Investigation, Qi Yan; Methodology, Qi Yan, Siqing Shan, Baishang Zhang and Menghan Sun; Project administration, Siqing Shan; Re-sources, Siqing Shan; Software, Qi Yan, Siqing Shan, Weize Sun and Yiting Luo; Supervision, Qi Yan and Siqing Shan; Validation, Qi Yan, Siqing Shan, Weize Sun and Menghan Sun; Visualization, Qi Yan, Baishang Zhang and Feng Zhao; Writing – original draft, Qi Yan and Menghan Sun; Writing – review & editing, Qi Yan, Siqing Shan, Baishang Zhang, Weize Sun, Menghan Sun, Feng Zhao and Xiaoshuang Guo.

Funding. This research was funded by National Natural Science Foundation of China, grant number 72071010 and by National Natural Science Foundation of China, grant number 71771010.

Competing interests. The authors declare no conflict of interest. The authors declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Ethical standards. This study was based on an analysis of social media data from Sina Weibo, which is a non-interventional study. We purchased the data collection service from Gooseeker. Gooseeker is an authorized API (Application Programming Interface) of Sina Weibo. The data does not involve private information such as personal name, gender, age, etc., and does not involve privacy and other related issues. All data can be used legally and does not require ethical approval.

Consent for publication. Not applicable.

References

1. **Potter CW.** A history of influenza. *J Appl Microbiol.* 2001;91:572-579. doi: 10.1046/j.1365-2672.2001.01492.x
2. **World Health Organization**. Global influenza strategy 2019-2030. Geneva. 2019. Licence: CC BY-NC-SA 3.0 IGO. Cataloguing-in-Publication (CIP) data. Accessed July 22, 2023. http://apps.who.int/iris
3. **Donnelly CA, Ghani AC, Leung GM, *et al.*** Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet.* 2003;361(9371):1761-1766. doi: 10.1016/s0140-6736(03)13410-1
4. **Domínguez-Cherit G, Lapinsky SE, Macias AE, *et al.*** Critically ill patients with 2009 influenza a(H1N1) in Mexico. *JAMA.* 2009;302(17):1880-1887. doi: 10.1001/jama.2009.1536
5. **Jia JS, Lu X, Yuan Y, *et al.*** Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature.* 2020;582(7812):389-394. doi: 10.1038/s41586-020-2284-y
6. **Hershfield HE, Scheibe S, Sims TL, *et al.*** When feeling bad can be good: mixed emotions benefit physical health across adulthood. *Soc Psychol Personal Sci.* 2013;4(1):54-61. doi: 10.1177/1948550612444616
7. **Karademas EC, Dimitraki G, Papastefanakis E, *et al.*** Emotion regulation contributes to the well-being of patients with autoimmune diseases through

illness-related emotions: a prospective study. *J Health Psychol.* 2020;25 (13-14):2096-2105. doi: 10.1177/1359105318787010
8. **Tyra AT, Griffin SM, Fergus TA, *et al.*** Individual differences in emotion regulation prospectively predict early COVID-19 related acute stress. *J Anxiety Disord.* 2021;81:102411. doi: 10.1016/j.janxdis.2021.102411
9. **Edo-Osagie O, De La Iglesia B, Lake I, *et al.*** A scoping review of the use of Twitter for public health research. *Comput Biol Med.* 2020;122:103770. doi: 10.1016/j.compbiomed.2020.103770
10. **Nguyen T, Larsen M, O'Dea B, *et al.*** Using spatiotemporal distribution of geocoded Twitter data to predict US county-level health indices. *Future Gener Comput Syst.* 2020;110:620-628, doi: 10.1016/j.future.2018.01.014
11. **Charles-Smith LE, Reynolds TL, Cameron MA, *et al.*** Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PLoS One.* 2015;10(10):e0139701. doi: 10.1371/journal.pone.0139701
12. **Shan S, Lin X.** Research on emergency dissemination models for social media based on information entropy. *Enterp Inf Syst.* 2018;12:888-909. doi: 10.1080/17517575.2017.1293300
13. **Alessa A, Faezipour M.** A review of influenza detection and prediction through social networking sites. *Theor Biol Med Model.* 2018;15(1):2, doi: 10.1186/s12976-017-0074-5
14. **Scanfeld D, Scanfeld V, Larson EL.** Dissemination of health information through social networks: Twitter and antibiotics. *Am J Infect Control.* 2010;38(3):182-188. doi: 10.1016/j.ajic.2009.11.004
15. **Aramaki E, Sachiko M, Mizuki M.** Twitter catches the flu: detecting influenza epidemics using factuality analysis. Proceedings of the 2011 Conference on empirical methods in natural language processing, Edinburgh, Scotland, UK, July 2011.
16. **Ahmed W, Bath PA, Sbaffi L, *et al.*** Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data. *Heatlth Info Libr J.* 2019;36(1):60-72. doi: 10.1111/hir.12247
17. **Yousefinaghani S, Dara R, Poljak Z, *et al.*** The assessment of Twitter's potential for outbreak detection: avian influenza case study. *Sci Rep.* 2019;9(1):18147. doi: 10.1038/s41598-019-54388-4
18. **Paul MJ, Dredze M, Broniatowski D.** Twitter improves influenza forecasting. *PLoS Curr.* 2014;6:ecurrents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117
19. **Aiello AE, Renson A, Zivich PN.** Social media- and internet-based disease surveillance for public health. *Annu Rev Public Health.* 2020;41:101-118. doi: 10.1146/annurev-publhealth-040119-094402
20. **Lampos V, Miller AC, Crossan S, *et al.*** Advances in nowcasting influenza-like illness rates using search query logs. *Sci Rep.* 2015;5:12760. doi: 10.1038/srep12760
21. **Masri S, Jia J, Li C, *et al.*** Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC Public Health.* 2019;19(1):761. doi: 10.1186/s12889-019-7103-8
22. **Samaras L, García-Barriocanal E, Sicilia M-A.** Comparing social media and Google to detect and predict severe epidemics. *Sci Rep.* 2020;10(1):4747. doi: 10.1038/s41598-020-61686-9
23. **Hoebel J, Maske UE, Zeeb H, *et al.*** Social inequalities and depressive symptoms in adults: the role of objective and subjective socioeconomic status. *PLoS One.* 2017;12(1):e0169764. doi: 10.1371/journal.pone.0169764
24. **Huynh VW, Chiang JJ.** Subjective social status and adolescent health: the role of stress and sleep. *Youth Soc.* 2018;50(7):926-946. doi: 10.1177/0044118X16646028
25. **Stringhini S, Carmeli C, Jokela M, *et al.*** Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet.* 2017;389(10075):1229-1237.
26. **Fournier MA.** Dimensions of human hierarchy as determinants of health and happiness. *Curr Opin Psychol.* 2020;33:110-114. doi: 10.1016/j.copsyc.2019.07.014
27. **Uecker JE, Wilkinson LR.** College selectivity, subjective social status, and mental health in young adulthood. *Soc Ment Health.* 2020;10(3):257-275. doi: 10.1177/2156869319869401
28. **Euteneuer F, Schäfer SJ, Neubert M, *et al.*** Subjective social status and health-related quality of life—a cross-lagged panel analysis. *Health Psychol.* 2021:40(1):71-76. doi: 10.1037/hea0001051
29. **McMaughan DJ, Oloruntoba O, Smith ML.** Socioeconomic status and access to healthcare: interrelated drivers for healthy aging. *Front Public Health.* 2020;8:231. doi: 10.3389/fpubh.2020.00231
30. **Kivimäki M, Batty GD, Pentti J, *et al.*** Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. *Lancet Public Health.* 2020;5(3): e140-e149. doi: 10.1016/S2468-2667(19)30248-8
31. **Wanberg CR, Csillag B, Douglass RP, *et al.*** Socioeconomic status and well-being during COVID-19: a resource-based examination. *J Appl Psychol.* 2020;105(12):1382. doi: 10.1037/apl0000831
32. **Peverill M, Dirks MA, Narvaja T, *et al.*** Socioeconomic status and child psychopathology in the United States: a meta-analysis of population-based studies. *Clin Psychol Rev.* 2021;83:101933. doi: 10.1016/j.cpr.2020.101933
33. **Tung J, Barreiro LB, Johnson ZP, *et al.*** Social environment is associated with gene regulatory variation in the rhesus macaque immune system. *Proc Natl Acad Sci U S A.* 2012;109(17):6490-6495. doi: 10.1073/pnas.1202734109
34. **Hassan Zadeh A, Zolbanin HM, Sharda R, *et al.*** Social media for nowcasting flu activity: spatio-temporal big data analysis. *Inf Syst Front.* 2019;21:743-760. doi: 10.1007/s10796-018-9893-0
35. **Perez-Rodríguez G, Pérez-Pérez M, Fdez-Riverola F, *et al.*** Mining the sociome for health informatics: analysis of therapeutic lifestyle adherence of diabetic patients in Twitter. *Futur Gener Comp Syst.* 2020;110:214-232. doi: 10.1016/j.future.2020.04.025
36. **Murayama T, Shimizu N, Fujita S, *et al.*** Predicting regional influenza epidemics with uncertainty estimation using commuting data in Japan. *PLoS One.* 2021;16(4):e0250417. doi: 10.1371/journal.pone.0250417
37. **Qin Z, Ronchieri E.** Exploring pandemics events on Twitter by using sentiment analysis and topic modelling. *Appl Sci (Basel).* 2022;12(23):21. doi: 10.3390/app122311924.
38. **Wang H, Xiong L, Wang C, *et al.*** Understanding Chinese mobile social media users' communication behaviors during public health emergencies. *J Risk Res.* 2022;25(7):874-891. doi: 10.1080/13669877.2022.2049621
39. **Sapolsky RM.** Social status and health in humans and other animals. *Annu Rev Anthropol.* 2004;33:393-418. doi: 10.1146/annurev.anthro.33.070203.144000
40. **Okamoto J, Johnson CA, Leventhal A, *et al.*** Social network status and depression among adolescents: an examination of social network influences and depressive symptoms in a Chinese sample. *Res Hum Dev.* 2011;8:67-88. doi: 10.1080/15427609.2011.549711
41. **Chang AC.** Chapter 5 - Machine and deep learning. In: Chang AC, ed. *Intelligence-Based Medicine*. Academic Press; 2020:67-140. doi: 10.1016/B978-0-12-823337-5.00005-6
42. **Lee C, Lee GG.** Information gain and divergence-based feature selection for machine learning-based text categorization. *Inf Process Manag.* 2006;42:155-165. doi: 10.1016/j.ipm.2004.08.006
43. **Jones KS.** Index term weighting. *Inf Storage Retrieva.* 1973;9(11):619-633. doi10.1016/0020-0271(73)90043-0
44. **Aizawa A.** An information-theoretic perspective of tf-idf measures. *Inf Process Manag.* 2003;39:45-65, doi: 10.1016/s0306-4573(02)00021-3
45. **Santos JC, Matos S.** Analysing Twitter and web queries for flu trend prediction. *Theor Biol Med Model.* 2014;11:11. doi: 10.1186/1742-4682-11-s1-s6
46. **Shan SQ, Yan Q, Wei YG.** Infectious or recovered? Optimizing the infectious disease detection process for epidemic control and prevention based on social media. *Int J Environ Res Public Health.* 2020;17(18):25. doi: 10.3390/ijerph17186853
47. **Uddin S, Hossain L, Wigand RT.** New direction in degree centrality measure: towards a time-variant approach. *Int J Inf Technol Decis Mak.* 2014;13:865-878. doi: 10.1142/s0219622014500217
48. **Gomez D, Gonzalez-Aranguena E, Manuel C, *et al.*** Centrality and power in social networks: a game theoretic approach. *Math Soc Sci.* 2003;46(1): 27-54. doi: 10.1016/s0165-4896(03)00028-3
49. **Alshahrani M, Zhu F, Sameh A, *et al.*** Efficient algorithms based on centrality measures for identification of top-K influential users in social networks. *Inf Sci.* 2020;527:88-107. doi: 10.1016/j.ins.2020.03.060
50. **Lee J, Lee Y, Oh SM, *et al.*** Betweenness centrality of teams in social networks. *Chaos.* 2021. doi: 10.1063/5.0056683

51. **Freeman LC.** Centrality in social networks' conceptual clarification. *Soc Netw.* 1979;1:215-239. doi: 10.1016/0378-8733(78)90021-7
52. **Carboni JL.** Social network analysis: methods and applications. *J Publ Adm Res Theory.* 2015;25:981-987. doi: 10.1093/jopart/muu083
53. **Cadini F, Zio E, Petrescu CA.** Using centrality measures to rank the importance of the components of a complex network infrastructure. In: Setola R, Geretshuber S, eds. *Critical Information Infrastructures Security*. Springer-Verlag; 2009:155-167.
54. **Likert R, Roslow S, Murphy G.** A simple and reliable method of scoring the Thurstone Attitude Scales. *Pers Psychol.* 2006;46(3):689-690. doi: 10.1111/j.1744-6570.1993.tb00893.x
55. **Xiang L.** *Recommendation System Practice*. Posts and Telecom Press; 2012:238-240.
56. **Spearman C.** The proof and measurement of association between two things. By C. Spearman, 1904. *Am J Psychol.* 1987;100(3-4):441-471. doi: 10.2307/1422689
57. **Heinen A, Valdesogo A.** Spearman rank correlation of the bivariate Student t and scale mixtures of normal distributions. *J Multivar Anal.* 2020;179:11. doi: 10.1016/j.jmva.2020.104650
58. **Mikolov T, Chen K, Corrado G, *et al.*** Efficient estimation of word representations in vector space. 2013:arXiv preprint. doi: 10.48550/arXiv.1301.3781
59. **Xu Y, Wang C, Dan Z, *et al.*** Deep recurrent neural network and data filtering for rumor detection on Sina Weibo. *Symmetry (Basel).* 2019;11:11. doi: 10.3390/sym11111408
60. **Conway M, Doan S, Kawazoe A, *et al.*** Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inform.* 2009;78(12):E47-E58. doi: 10.1016/j.ijmedinf.2009.03.010
61. **Metz CE.** Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8:283-298. doi: 10.1016/s0001-2998(78)80014-2
62. **Aslan S.** A novel TCNN-Bi-LSTM deep learning model for predicting sentiments of tweets about COVID-19 vaccines. *Concurr Comput.* 2022;34(28):e7387. doi: 10.1002/cpe.7387
63. **Narra M, Umer M, Sadiq S, *et al.*** Selective feature sets based fake news detection for COVID-19 to manage infodemic. *IEEE Access.* 2022;10:98724-98736. doi: 10.1109/access.2022.3206963
64. **Raj C, Meel P.** ARCNN framework for multimodal infodemic detection. *Neural Netw.* 2022;146:36-68. doi: 10.1016/j.neunet.2021.11.006
65. **Shan SQ, Zhao F, Wei YG, *et al.*** Disaster management 2.0: a real-time disaster damage assessment model based on mobile social media data: a case study of Weibo (Chinese Twitter). *Saf Sci.* 2019;115:393-413. doi: 10.1016/j.ssci.2019.02.029
66. **Chen SJ, Mao J, Li G, *et al.*** Uncovering sentiment and retweet patterns of disaster-related tweets from a spatiotemporal perspective - a case study of Hurricane Harvey. *Telemat Inform.* 2020;47:18. doi: 10.1016/j.tele.2019.101326
67. **Woo H, Cho Y, Shim E, *et al.*** Estimating influenza outbreaks using both search engine query data and social media data in South Korea. *J Med Internet Res.* 2016;18:11. doi: 10.2196/jmir.4955
68. **Gao YZ, Wang SW, Padmanabhan A, *et al.*** Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *Int J Geogr Inf Sci.* 2018;32:425-449. doi: 10.1080/13658816.2017.1406943
69. **Yoo S, Kim D, Yang SM, *et al.*** Real-time disease detection and analysis system using social media contents. *Int J Web Grid Serv.* 2020;16:22-38.
70. **Ma JW, Yang Y, Wilson JAJ.** A window to the ideal self: a study of UK Twitter and Chinese Sina Weibo selfie-takers and the implications for marketers. *J Bus Res.* 2017;74:139-142. doi: 10.1016/j.jbusres.2016.10.025

https://doi.org/10.1017/dmp.2023.117 Published online by Cambridge University Press