International Journal of
Environmental Research and Public Health

MDPI

Article

Infectious or Recovered? Optimizing the Infectious Disease Detection Process for Epidemic Control and Prevention Based on Social Media

Siqing Shan [1,2], Qi Yan [1,2,*] and Yigang Wei [1,2]

[1] School of Economics and Management, Beihang University, Beijing 100191, China; shansiqing@buaa.edu.cn (S.S.); weiyg@buaa.edu.cn (Y.W.)
[2] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operation, Beijing 100191, China
* Correspondence: bhyanqi@buaa.edu.cn

Received: 7 August 2020; Accepted: 16 September 2020; Published: 19 September 2020

Abstract: Detecting the period of a disease is of great importance to building information management capacity in disease control and prevention. This paper aims to optimize the disease surveillance process by further identifying the infectious or recovered period of flu cases through social media. Specifically, this paper explores the potential of using public sentiment to detect flu periods at word level. At text level, we constructed a deep learning method to classify the flu period and improve the classification result with sentiment polarity. Three important findings are revealed. Firstly, bloggers in different periods express significantly different sentiments. Blogger sentiments in the recovered period are more positive than in the infectious period when measured by the interclass distance. Secondly, the optimized disease detection process can substantially improve the classification accuracy of flu periods from 0.876 to 0.926. Thirdly, our experimental results confirm that sentiment classification plays a crucial role in accuracy improvement. Precise identification of disease periods enhances the channels for the disease surveillance processes. Therefore, a disease outbreak can be predicted credibly when a larger population is monitored. The research method proposed in our work also provides decision making reference for proactive and effective epidemic control and prevention in real time.

Keywords: social media; flu; disease detection; sentiment analysis; text classification

1. Introduction

The traditional infectious disease detection process is being challenged by potential social media applications [1,2]. The latest estimates released by the United States Centers for Disease Control and Prevention (US-CDC) revealed the worldwide severity of the illness. According to this authoritative report, the US-CDC estimates that in the period between 1 October 2018 and 4 May 2019, there were approximately 37.4 million to 42.9 million flu infectious in the population population, among which there were from 17.3 to 20.1 million flu-related medical visits [3]. Furthermore, 531,000 to 647,000 people require flu-related hospitalizations, and unfortunately, influenza caused 36,400–61,200 estimated deaths. This estimate is based on more recent data from a larger and more diverse group of countries, including lower middle-income countries, and this estimate excludes deaths from non-respiratory diseases. The statistics indicate a severe reality and a pressing challenge because influenza causes significant losses of human life and damage to property worldwide. As evidenced by the current flu season, influenza viruses can rapidly mutate, evading the most current vaccine formulations [4]. Infectious diseases continue to be the leading cause of death worldwide, and they cause serious loss of life and property when they cannot be quickly and accurately assessed [5]. The World Health

Organization (WHO) announced on 15 August 2003 that by 7 August 2003, there were 8422 infectious cases of Severe Acute Respiratory Syndrome (SARS) worldwide, which involved 32 countries and regions [6]. The number of deaths due to SARS in the world totaled 919, and the mortality rate was almost 11%. A handful of studies estimated the global macroeconomic impact of SARS at USD 30–100 billion, or approximately USD 3–10 million per case [7]. However, from the first reported case on 15 December 2002 to the first epidemic announcement by the Chinese government on 11 February 2003, the time span was almost two months. The consequence was a significant loss of life and property due to the lack of effective disease identification and monitoring methods for real-time epidemic control. Based on the infectious disease epidemic report issued by the National Health Commission of China, in recent years, the number of incidences and deaths due to influenza in C-class infectious diseases shows an explosively upward trend annually (see Figure 1). The death toll for five months during 2019 was nearly two times that for all of 2018.

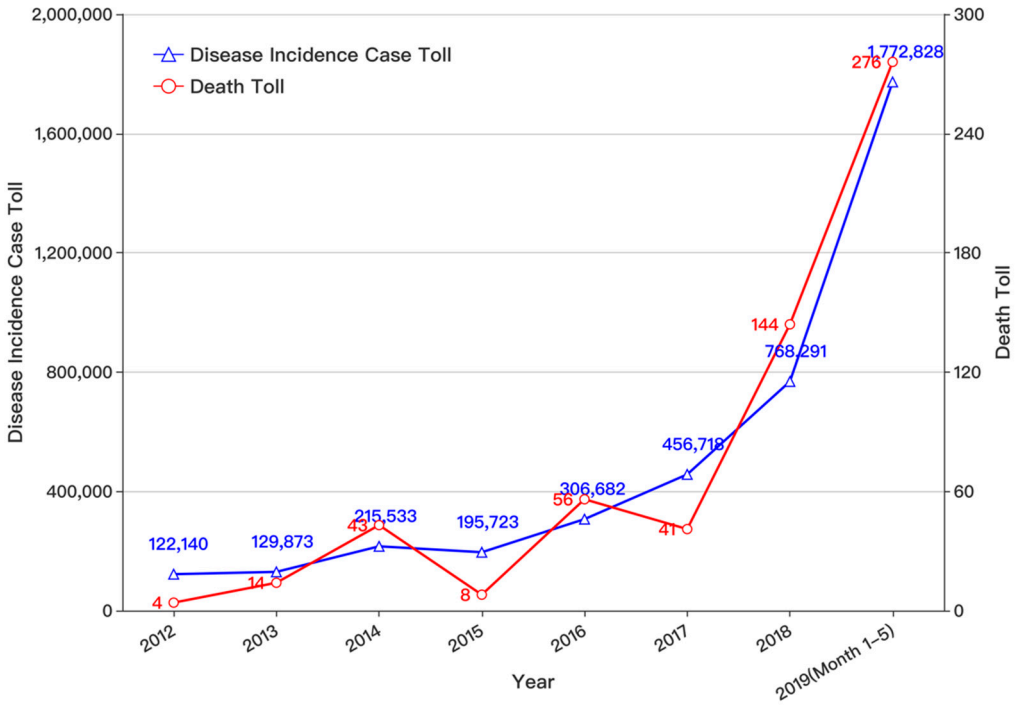


Figure 1. Number of People in the Incidences and Deaths from Flu in China.

The fundamental cause of these serious outbreaks is summarized as follows. First, traditional disease prevention and control institutions mainly rely on a single channel for information monitoring and access. Specifically, data are exclusively sourced from clinical statistics. However, relying exclusively on clinical statistics has obvious disadvantages, such as being time consuming and creating high labor costs [8,9]. The traditional detection methods cannot integrate multichannel infectious disease information, such as social media and search engine data. Second, social media data are a powerful and promising tool that have been applied to many research subjects, such as healthcare informatics [10,11], sentiment analytics [12], and disaster management [13–15]. The remarkable value of social media has been widely recognized [16]. Particularly in this paper, we refer to microblogs posted on social media platforms especially from Sina Weibo as weibos. Although disease prevention and control institutions have exerted a significant role in disease detection, they might come up against substantial difficulties in using social media data in disease detection and control due to the lack of analytic methods or accurate monitoring of outbreaks and periods of infectious diseases. There is still a long journey to go to fully utilize social media data in disease prevention and control. Third, the flu is characterized by strong contagiousness and rapid spread, which makes it difficult to monitor in real-time or precisely estimate the spread of the flu. For example, the flu can be easily spread

by droplets or contaminated items in the air and contact among people. The Centers for Disease Control and Prevention (CDC) publishes data on influenza-like illness ILI based on statistics and evaluation after a patient's visit. It would be extremely difficult to obtain information and perform an analysis before the visit. However, apparently, time lags could lead to delayed treatment. Fourth, Google Flu Trends (GFT) provided an estimate of more than double the proportion of clinical data for influenza-like illness (ILI) published by the Centers for Disease Control and Prevention (CDC) [17]. The ILI was calculated based on surveillance reports from laboratories across the United States [18]. Google used web search data to propose a GFT model for real-time monitoring. When patients are aware of their active flu period and search for flu-related keywords through search engines such as Google, the patients' behavior is recorded by the search engine. Social media sensors, in contrast, show unique advantages for quick flu monitoring and reliable estimation and prediction. There has been a growing consensus that social media sensors can also perform real-time monitoring that is more accurate than GFT [8]. Fifth, part of the ILI data cannot be collected and thus are not available if the patient does not go to the hospital, which makes disease monitoring information inaccurate and incomplete. Therefore, the severity and urgency of the disease as reflected by traditional statistics are often underestimated. However, social media can capture this part of the data because when bloggers catch a cold, they can post their own flu symptoms through social media. If bloggers do not realize that they catch flu, they can also post some flu-related microblogs on social media. This part of the data can also be seen in advance of CDC released reports [19]. Sixth, the previous literature mainly focused on whether the bloggers are infected based on the social media platform [20] and cannot accurately distinguish the various periods of the disease, which largely reduces the effectiveness and pertinence of the information for disease control measures.

This section provides an extensive review of the relevant literature in three parts: optimization of the infectious disease detecting process, social media utilization for disease detection and semantic analysis techniques based on social media data. Much of the current research focuses on social media to analyze short texts [21–23], in which several papers predict flu trends by classifying flu-related social media data [19]. However, these studies do not divide the flu periods any further. Speedily evolving infectious diseases, including SARS, Ebola and influenza, pose significant health threats throughout the world because of their rapidly changing status and complicated detection process [24,25]. A considerable amount of work is devoted to forecasting disease outbreaks. A two-period model optimized the process of when and where to assign Ebola treatment units across geographic regions during the outbreak's early phases [26]. Chen et al. (2018) developed a mixed-integer programming (MIP)-based framework to systematically analyze a rich set of policies and to determine the optimal hepatocellular carcinoma surveillance policies that maximize the societal net benefit [27]. It is observed that the surveillance policies should be adapted to different disease progression rates and states. Several studies in this area focused on finding optimal surveillance solutions for flu vaccine production and allocation [28,29]. Another study considered the conditions of limited reporting and spatial aggregation on the optimization of influenza surveillance system design [30]. Research in this area is mainly concerned with epidemiological inference and prediction based on clinical data collection, but few of the studies provide improved detecting measures using social media data, which represent an alternative source of passive traditional surveillance data that have a larger volume and fewer reporting delays.

Data from social networks show apparent advantages in several aspects, such as being real time and time-sharing, along with a broad scope of data coverage [31–34]. Disease surveillance has been investigated based on the recent rise in the popularity and scale of social media data. Aiello et al. (2011) reviewed and addressed the use, promise, perils, and ethics of social media- and internet-based data collection for public health surveillance [35]. Additionally, the infectious disease detection process is being challenged by the potential applications of social media [1,2,36]. Multiple types of social media have become emerging and promising data sources of disease surveillance and shown advanced achievements in tracking health informatics in different areas all over the world. Raamkumar et al.

(2020) examined the differences of COVID-19-related public responses on Facebook in the United States, England and Singapore and showed that social media analysis was capable of providing insights about the communication strategies during disease outbreaks [37]. Lwin et al. (2018) and Vijaykumar et al. (2017) investigates how Facebook can be utilized to implement and adapt in responding to the Zika epidemic in Singapore [38,39]. Moreover, Dubey et al. (2014) identified and evaluated YouTube as significant resource for providing and disseminating information on public health issues like West Nile virus infection [40]. Davidson et al. (2015) constructed an empirical network to substantially improve performance in predicting infections one week into the future using CDC data and combining this with internet-based data in the U.S. [41]. Chen et al. (2016) proposed two temporal topic models to capture hidden states of a flu-related user and get better flu-peak predictions by using Twitter data. In addition, they validated their approaches by modeling the flu using Twitter in multiple countries of South America [42]. Lamb et al. (2013) demonstrated that the use of Twitter data leads to significant improvements in flu surveillance by discriminating those categories of flu tweets that reported infection from those that expressed concerned awareness of the flu as well as tweets about the authors versus those about others [43].

Sentiment classification [44,45], feature extraction [46] and public opinion monitoring [47,48] are performed based on a social network dataset for sentiment analysis. Chen et al. (2020) introduced a novel approach of adding semantics as additional features into the training set for sentiment analysis, and they applied this approach to predict sentiment for three different Twitter datasets [49]. The authors also investigated the real-time flu detection problems and proposed a flu detection model with emotional factors and semantic information [50]. Adamopoulos et al. (2018) examined the effect of latent personality traits on consumers' behavior and preferences, which originated from social media users' levels of emotional range [51]. The effect of social media advertising content on customer engagement was also studied via Facebook users' humor and emotion [52]. Although sentiment analysis has been widely applied to many fields, few researchers consider it to be a powerful tool to be used in determining a patient's status when detecting infectious diseases. Furthermore, word embedding techniques, such as bag of words [53] and word2vec [54], have become effective means of text processing. High-quality word vector representations provide distributional information about words [55], especially for word2vector, which appears to be outstanding at improving a model's performance on a limited amount of data. The development of word2vector significantly improves the effectiveness of word representations by transforming sparse, discrete and high-dimensional vectors into dense, continuous and low-dimensional vectors [56]. It is a foundation to transform word segments into fixed dimensional vectors, namely, word embedding, when studying user-generated content [57,58]. In this paper, the use of high dimensional numeral vectors to represent words serve as semantic feature extractors or the input variables of a neural network. Artificial Neural Network (ANN) techniques have been used widely in text classification. Hughes et al. (2017) have used Convolutional Neural Networks (CNNs) for text processing and classification in online news, reviews and medical text [59]. A large number of modified ANN techniques have emerged with technical advancements. Recurrent Neural Networks (RNNs) have also been an effective method for speech recognition and text sequence tagging [60], and Long Short Term Memory (LSTM) networks [61] perform well for sequence-based learning tasks. In addition, a large amount of text processing research has emerged, including the use of part-of-speech tagging [62], lexicon approaches [63], and other deep learning techniques [64]. The extant literature has several key limitations. Firstly, the abovementioned studies have mainly focused on a single aspect of text processing, sentiment classification or flu tracking [43,62]. More importantly, these papers do not adequately consider the influence of sentiment polarity on the classification of flu-related weibos or on dividing the different periods of flu-related weibos [50]. Therefore, this paper incorporates sentiment factors into flu surveillance research, and the period classifications of flu-related weibos are probed. Second, neural network techniques are used to process the weibos at the text level [61,65]. LSTM networks can be used for sentiment analysis of film reviews, part-of-speech tagging, and other fields [60]. Previous studies show that LSTM performs relatively

well in text processing but has been rarely used for disease weibo analysis. To fill in this gap, this paper aims to investigate the relationship between sentiment polarity and the flu period at the word level and text level based on a weibo dataset.

The main research rationale of this study is straightforward, i.e., first, to investigate the relationship between the sentiment polarity and the flu period from social networks, and second, to optimize the disease detecting process by predicting the different periods of flu.

2. Materials and Methods

2.1. Research Model

The model proposed in this paper can detect infectious diseases through multiple channels; they can be perceived earlier and the model larger populations and more efficiency than traditional processes (See Figure 2).

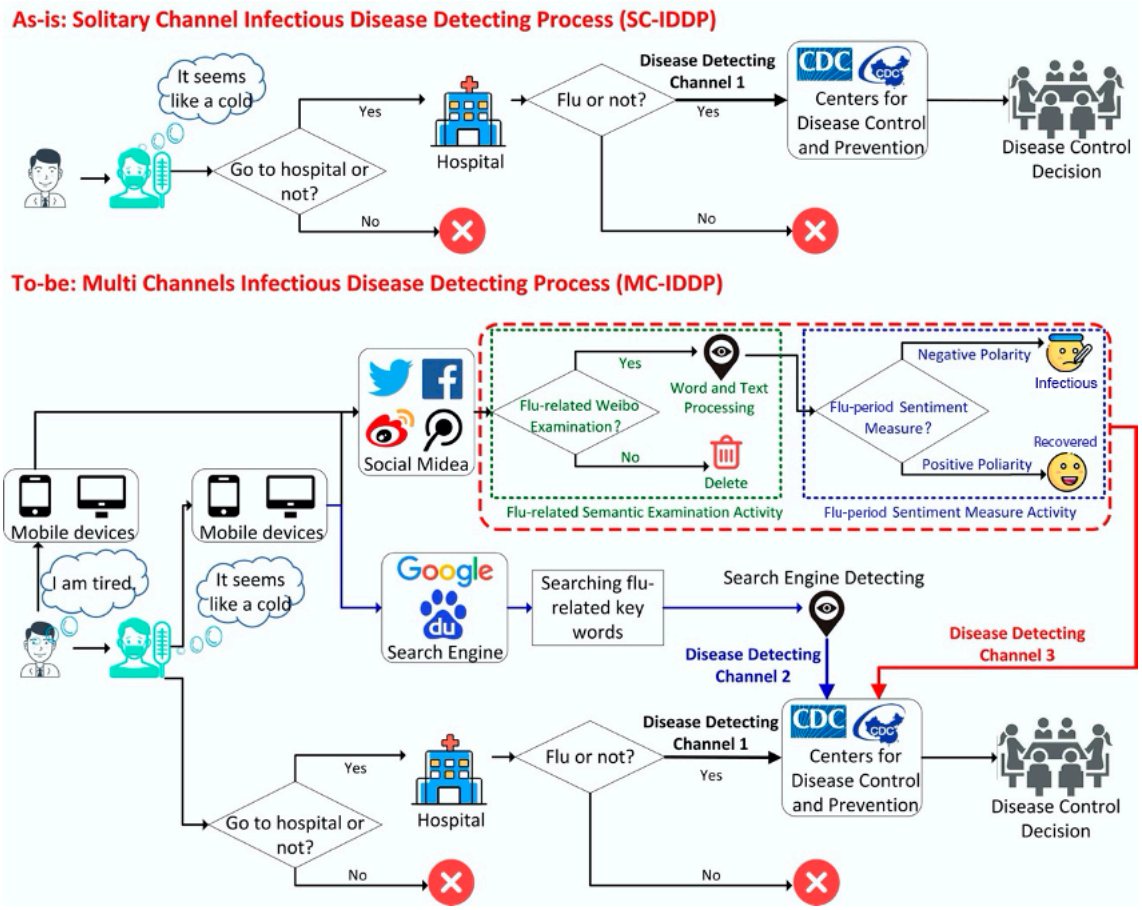


Figure 2. Solitary Channel and Multi-Channels Process.

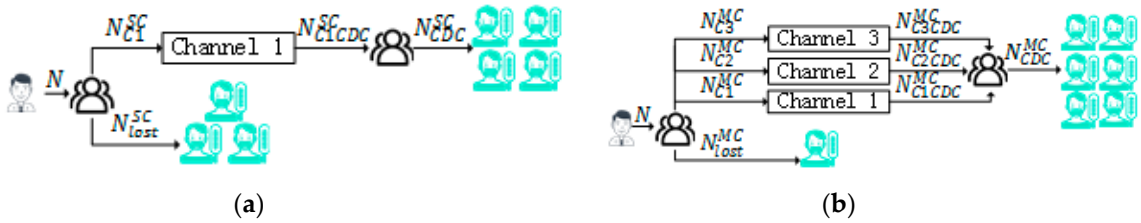
The upper part of Figure 2 is a schematic diagram of a traditional solitary infectious disease detecting process. Patients usually go to the hospital when they have already suffered influenza. Only afterwards it is possible for the CDC to detect the flu trends of the whole society. In the "as-is" process, there is only one channel to detect infectious diseases, which is through hospital institutions. The "as-is" process is abbreviated as SC-IDDP. The "to-be" process is the multiple channel infectious disease detection process (MC-IDDP). The bottom part of Figure 2 illustrates the MC-IDDP process based on social media platforms. MC-IDDP has three infectious disease detecting channels. Channel 1 is the traditional infectious disease detecting channel, mentioned above. Channel 2 monitors the infectious diseases by using search engines [17,18]. This paper constructs Channel 3 to detect infectious

diseases using social media data. Compared with Channel 1 and 2, Channel 3 enables more effective disease detection for the following three reasons. First, the traditional Channel 1 is not able to use social media data to detect diseases, such as flu, whereas Channel 2 and Channel 3 can use these data. Second, search engines can detect the keywords of the corresponding flu symptoms and treatments in Channel 2 when people realize that they have the flu. It neglects some of the patients who have caught the flu but do not search for flu keywords and simultaneously adds fake patients who run the search engine for flu keywords without flu. Third, although people do not realize that they have caught the flu, they usually post tweets via social media to reflect their feelings, opinions and behaviors, which can be detected by Channel 3. It helps the CDC to analyze social media data to discover early flu trends. Additionally, Channel 3 can not only monitor infectious disease outbreaks in early stages but also identify the flu period. The content inside the dotted line (shown in Figure 2) is the main research content of this paper, which aims to find the flu-related weibos and further determine the flu period to improve the accuracy of infectious disease detection.

As the "to-be" process optimized by social media, the MC-IDDP process can improve the reliability of detecting infectious disease and discover more infectious people. It recognizes the disease periods earlier and works in an efficient way. Next, we propose 4 propositions and corresponding demonstrations.

Proposition 1. *The MC-IDDP process detects a larger number of potentially infectious people in the population.*

Demonstration 1. *The logical structure model of MC-IDDP is shown in Figure 3. Figure 3a shows the logical structure model of a solitary channel, and Figure 3b shows the logical structure model for multiple channels.*



(**a**) (**b**)

Figure 3. The Logical Structure Model of a Solitary Channel and Multiple Channels. (**a**) Solitary Channel; (**b**) Multiple Channels.

The main parameters are described as follows. N represents the number of possible infectious people; N_{C1}^{SC} represents the number of possibly infectious people recorded by hospitals using a solitary channel; N_{lost}^{SC} represents the number of possibly infectious people but do not go to the hospital in a solitary channel; N_{C1CDC}^{SC} represents the number of infectious people who are diagnosed and reported to the CDC in a solitary channel; N_{C1}^{MC} represents the number of possibly infectious people recorded by hospital in multiple channels; N_{C1CDC}^{MC} represents the number of infectious people who are diagnosed and reported to the CDC in Channel 1; N_{C2}^{MC} represents the number of possible patients who search disease-related information through search engines in a multiple channel environment; N_{C2CDC}^{MC} represents the number of infectious people who are detected and reported to the CDC after analyzing the search data in Channel 2; N_{C3}^{MC} represents the number of possible patients who post related microblogs through social media in a multiple channel environment; N_{C3CDC}^{MC} represents the number of infectious people who are detected and reported to the CDC after analyzing social media data in Channel 3; N_{lost}^{MC} represents the number of possible patients who do no go to the hospital, and do not use search engines and social media in multiple channels; N_{CDC}^{SC} represents the number of infectious people who are diagnosed and reported to the CDC in a solitary channel; N_{CDC}^{MC} represents the number of infectious people who are detected by the three channels and reported to the CDC.

Obviously, in a solitary channel,

$$N = N_{C1}^{SC} + N_{lost}^{SC} \quad (1)$$

In multiple channels,

$$N = N_{C1}^{MC} \cup N_{C2}^{MC} \cup N_{C3}^{MC} + N_{lost}^{MC} \quad (2)$$

Without the loss of generality, it is assumed that the number of infectious people who are admitted to the hospital remains the same regardless of whether it is a solitary channel or multiple channels, which is,

$$N_{C1}^{MC} = N_{C1}^{SC} \quad (3)$$

The following relationship exists,

$$N_{C1}^{MC} \cup N_{C2}^{MC} \cup N_{C3}^{MC} > N_{C1}^{SC} \quad (4)$$

According to Formula (4), in multiple channels, more infectious people can be detected. Proposition 1 is validated.

Proposition 2. *The MC-IDDP process recognizes the patients early and detects the infectious disease in time.*

Demonstration 2. *The Google flu trends model provides a promising measure in that the search engine records the users' relevant search data on disease symptoms and treatments, to detect infectious diseases. Additionally, some of the patients do not realize that they might be infected but post weibos with disease symptoms. Therefore, people's sentiment and physical conditions can be reflected in social media, which is an important infectious disease sensor. We added time information in Figure 3b to generate Figure 4. The horizontal axis t indicates the earliest time at which each channel can detect infectious disease information.*

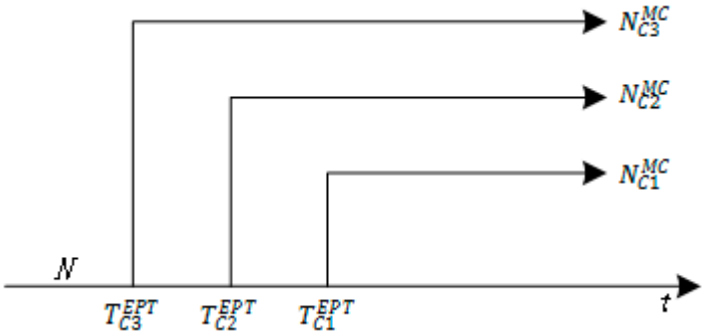


Figure 4. The Logical Structure Model over Time.

In Figure 4, EPT represents the Earliest Possible Time, and T_{Ci}^{EPT} represents the earliest possible time to detect the disease by Channel i. Obviously, we conclude that

$$T_{C3}^{EPT} < T_{C2}^{EPT} < T_{C1}^{EPT} \quad (5)$$

According to Formula (5), the EPT for Channel 3 is the smallest, and the EPT for Channel 1 is the largest. Therefore, the MC-IDDP process based on social media data can achieve more timely monitoring.

Proposition 3. *The MC-IDDP process can conduct infectious disease detection more efficiently.*

Demonstration 3. *In the MC-IDDP process, Channel 1 requires a large number of doctors and staff with high operating costs, but Channel 2 and Channel 3 rely only on big data and analytical tools to conduct the surveillance. Compared with Channel 1, the operation costs of Channel 2 and Channel 3 can be negligible. Therefore, the total cost of the MC-IDDP process is almost the same as the total cost of the SC-IDDP process. However, the MC-IDDP process can achieve a wider range of detection (according to Proposition 1), and thus, the MC-IDDP process has higher monitoring efficiency. Proposition 3 is validated. The Detection Accuracy is*

defined as the percentage of the correct number of patients out of the total number of possible patients who have been monitored through the infectious disease detection channels. In this paper, the Detection Accuracy (Acc_D) can be calculated by the following equation,

$$Acc_D = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP is an abbreviation of True Positive, which indicates that an infectious patient is diagnosed as a patient; TN is an abbreviation of True Negative, which means that a nonpatient is diagnosed as a nonpatient; FP is an abbreviation of False Positive, which indicates that a nonpatient is misdiagnosed as a patient (Type I error); FN is an abbreviation for False Negative, which means that a real patient is misdiagnosed as a nonpatient (Type II error).

Proposition 4. *The detection accuracy in any channel of the MC-IDDP process helps to improve the entire accuracy of the disease detection.*

Demonstration 4. *The MC-IDDP process has three disease detection channels. Each channel has independent methods and techniques. The detection accuracy of Channel 1 depends on the hospital's medical plan and medical technology. The detection accuracy of Channel 2 depends on the statistical analysis and technical means used by the search engine. The detection accuracy of Channel 3 depends on the semantic analysis and machine learning application to social media data. The detection results of the three channels do not affect one another, and thus, the detection accuracy of each channel is related to only the method and supporting technology of that channel. Increasing the detection accuracy in any channel can improve the entire accuracy of the disease detection. The proposition is validated.*

This paper focuses on detection Channel 3. To improve the detection accuracy of Channel 3 using social media data, this paper proposes an effective dual analytical activity model to determine the status and period of the infected population. The next section discusses the infectious disease detection Channel 3 and the main activities.

2.2. Infectious Disease Detection Channel 3

The Infectious Diseases Detection Channel 3 is a crucial channel for the MC-IDDP process that makes direct use of social media data. It has two main activities: flu-related semantic examination activity and flu-period sentiment measure activity, and thus, Channel 3 is also named the dual analytical activity model, as shown in Figure 5. The contents and features of the model are described in detail in the following section.

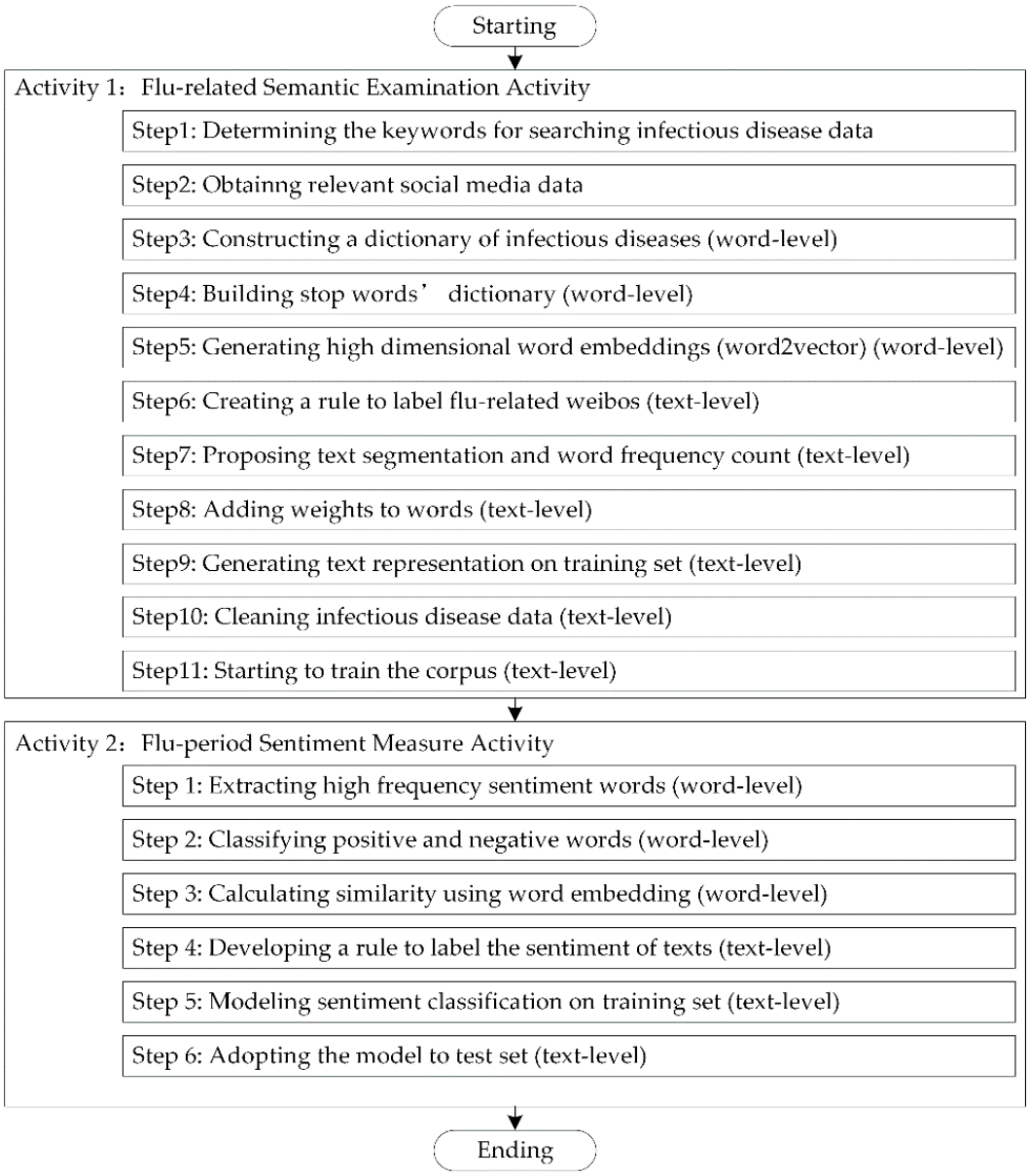


Figure 5. Dual Analytical Activity Model in Channel 3.

2.2.1. Flu-Related Semantic Examination Activity

The purpose of the Flu-related semantic examination activity is to enable semantic recognition and analysis of social media data based on infectious diseases (taking the flu as an example). The main content of the activity includes at least 12 steps, such as determining keywords for searching infectious disease data and obtaining relevant social media data, as shown in Figure 5. Next, we provide some crucial steps.

1. Obtaining Flu Data Based on Social Media

Six keywords were crawled from Sina Weibo; these six keywords include "flu (Gan Mao)", "influenza (Liu Gan)", "cough (Ke Sou)", "fever (Fa Shao)", "sneeze (Pen Ti)" and "nasal congestion (Bi Sai)". We used Python for this task, and these flu-related weibos constitute an elementary corpus.

2. Cleaning the Flu Data

Considering the existence of advertisements and forwarding, this study used Support Vector Machine (SVM) to screen the invalid data and retain the flu-related weibos. The final flu-related weibo

corpus is generated after word segmentation and the removal of stop words. The entire process is shown in Figure 6.

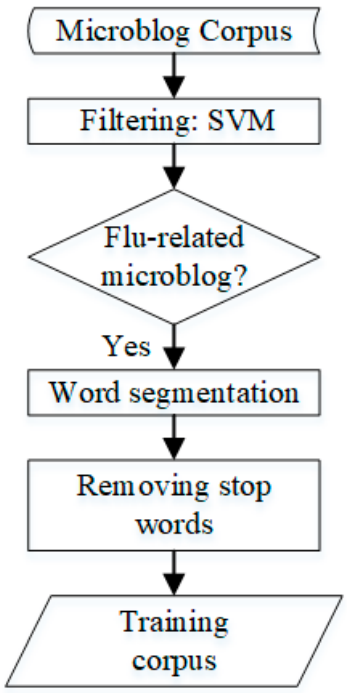


Figure 6. Cleaning Data.

2.2.2. Flu-Period Sentiment Measure Activity

This activity is mainly to accurately identify the patient's disease period to improve the detection accuracy of infectious diseases. The whole process includes six steps. Activity 2 in Figure 5 shows that the word level detection consists of 3 steps and the text level includes 3 steps, which is introduced as follows.

1. Relationships between Sentiment Polarity and Flu Period at the Word Level

Word2vector is an efficient training method to transform a symbol into a structure and digital representation. Word embedding is represented differently in different vocabularies or by different training methods. The principles of Word2vector are mainly separated into three parts to transform the words from the vocabulary of a flu-related weibo into high-dimensional space vectors, as follows in Figure 7.

- Building the vocabulary of the flu-related weibo texts: the processing of the text, which means that a specific vocabulary is required;
- Initializing the network structure of the weibo text: the initialization of parameters in the CBOW model, with Huffman coding generation;
- Saving the word embedding: saving the result in a specific form.

(CBOW Model) The CBOW model contains three layers, including the input layer, projection layer and output layer. The window size used in this paper is 5. The vectors that correspond to each word are first found to be summarized from the input layer to the projection layer. After all of the word vectors in the window are gathered, they are stored in the projection layer, and the mean value is calculated.

(Hierarchical Softmax) Hierarchical Softmax is a key technology that is used in word2vec to improve the performance. In the Huffman tree, the softmax mapping of the hidden layer to the output layer is proceeded step by step along the Hoffman tree and, thus, this softmax is named "Hierarchical Softmax".

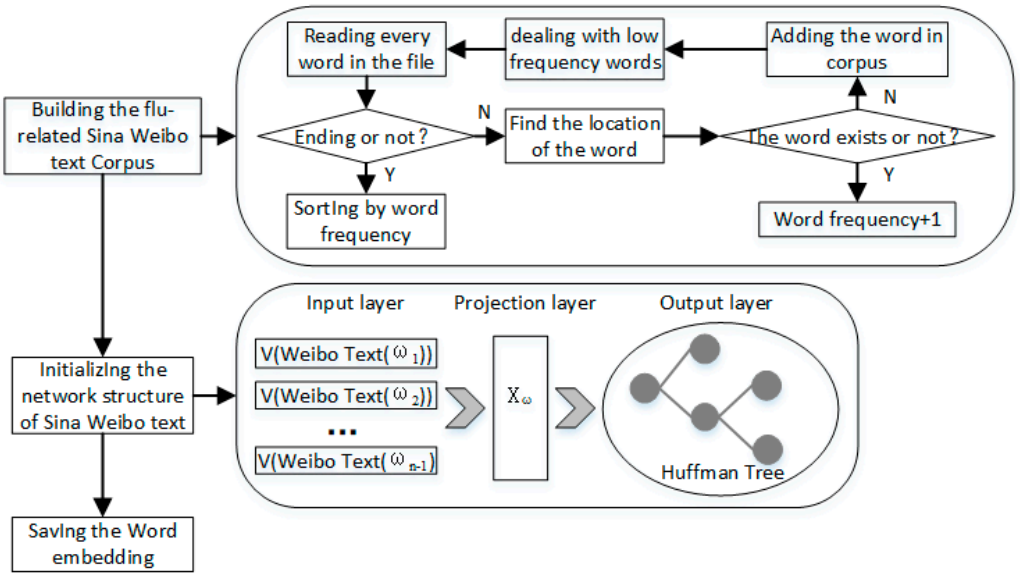


Figure 7. Word2vector Principles of Sina Weibo Text.

2. Detecting the Flu Period Based on Sentiment Polarity at the Text Level

Recurrent Neural Networks (RNNs) perform well in text classification. However, long-term dependence occurs if the interval of two words is overly large. This paper adopts a novel type of RNN called Long Short-Term Memory that works better than traditional RNNs on tasks that involve long time lags. Its architecture permits LSTM to bridge massive time lags between relevant input events (1000 steps and more) [65]. Figure 8 shows the structure of the network with 8 main layers, and we describe each layer below.

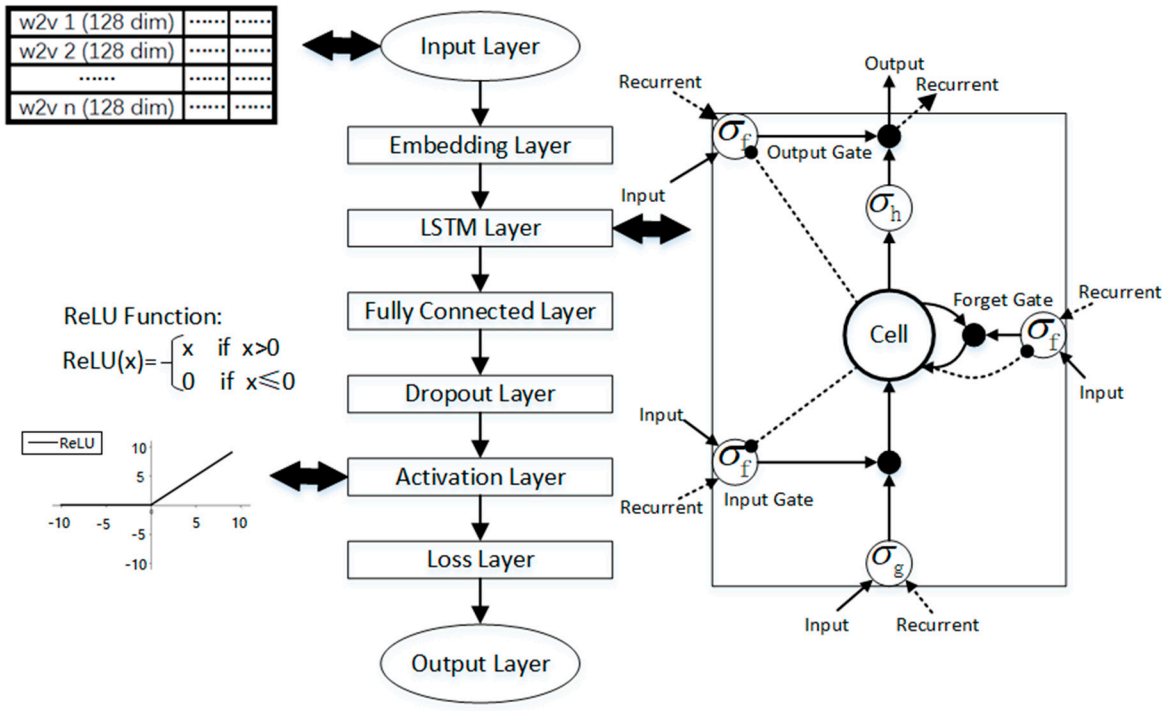


Figure 8. Structure of the Long Short Term Memory Networks Constructed by Layers.

Two LSTM neural networks in parallel are built to perform binary classification for sentiment classification and flu-period classification in this section. The input is the word embedding trained by the word2vector based on the corpus of flu-related weibos. The entire process is shown in Figure 9.

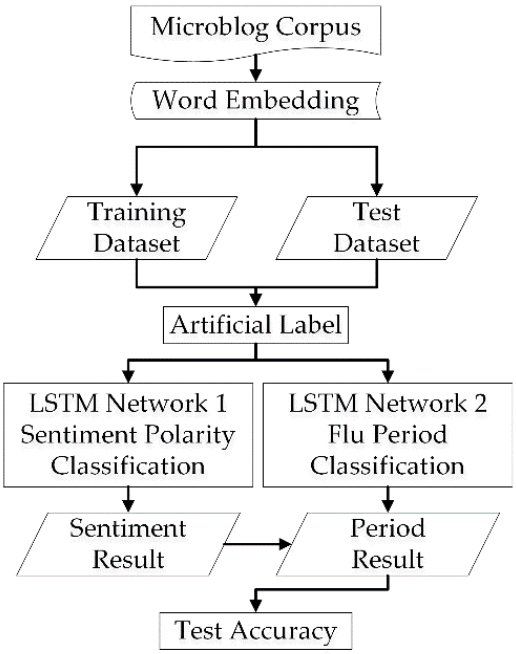


Figure 9. Process of Flu Period and Sentiment Classification.

2.2.3. Construction of LSTM for Sentiment Polarity and Flu Period Classification

The LSTM network is used to construct two neural networks to classify flu-related weibos. The first is to classify the sentiment polarity. The other intends to classify the period of the flu bloggers. Each neural network consists of 8 main layers, as follows.

(Input layer and masking layer) The first layer is the input layer, which uses 128-dimension vectors by means of the word2vector algorithm. The mask value is set to 0 in the masking layer.

(LSTM layer) Each LSTM unit is a storage unit that controls the passage or filtering of information through the three gates, to alter the cell state. "Implementation" is set to 2 to combine the input gate, forget gate, and output gate into a single matrix for more efficient operations.

(Fully connected layer) The core operation of the full connection is the matrix vector product. The essence is a linear transformation from one feature space to another feature space. A dense layer is used.

(Dropout layer) The settings of this layer are mainly to prevent overfitting in neural network training. The parameters of this layer are set to 0.3, which implies that the unit that was transferred from the LSTM layer will be randomly discarded by 30% during training, leaving 70% of the units used.

(Activation layer) The activation function in this layer is "ReLU" (Rectified Linear Units), and as a result, the convergence rate of the model is maintained at a steady state.

(Loss layer and output layer) Since the classification of the sentiment polarity and the period are all binary classifications, binary cross entropy is agreeable as a loss function, with the accuracy rate as a metric of the model. The classification result and test score are received through the output layer.

2.3. Data Description

According to the 41st China Statistical Report on Internet Development published by the China Internet Network Information Center (CNNIC), Weibo has 316.01 million users and a user usage rate of 40.9% on social media by December 2017 [66]. Sina Weibo has maintained the top rank in China's

weibo. According to the second quarter earnings released by Sina, monthly active users from Sina Weibo reached 431 million by 30 June 2018, outstripping Twitter as the world's largest independent social media company in terms of user scale [67]. Sina Weibo has always been the data source of various types of major events and emergencies in China and has a far-reaching scope of dissemination and important social influence. This study uses web-based social media data in Sina Weibo. The details are clearly shown in Table 1.

Table 1. Data Description.

Category	Field Name	
Tweet's information	URL, released time, title, text	
Blogger's Information	blogger's ID, nickname of the blogger	
Resource	Sina Weibo	
Keywords	flu (Gan Mao), influenza (Liu Gan), cough (Ke Sou), fever (Fa Shao), sneeze (Pen Ti), nasal congestion (Bi Sai)	
Amount of word2vector training corpora	In 2016 In 2017	50,000 50,000
Amount of labeling sets	In 2016 In 2017	10,000 10,000
Total valid amount	15,301	
LSTM training set	10,711	
LSTM test set	4590	

We collected the texts that contained the keywords, namely, "flu (Gan Mao)", "influenza (Liu Gan)", "cough (Ke Sou)", "fever (Fa Shao)", "sneeze (Pen Ti)" and "nasal congestion (Bi Sai)" in 2016 and 2017 through a Sina Application Programming Interface (API). We purchased the official data collection service from Gooseeker. Gooseeker is an authorized API of Sina Weibo. The data do not include private information such as personal name, gender, age, etc., and do not endanger privacy and other related issues. All data can be used legally. However, the data include a large amount of advertising and unrelated material. Thus, the Support Vector Machine (SVM) is used to filter out unrelated flu weibos. Ultimately, 100,000 flu-related weibos were chosen randomly as a word2vector training corpus in which 10,000 weibos in 2016 and 10,000 weibos in 2017 were randomly selected as a neural network classification dataset of sentiment polarity and period. In what follows, the 20,000 weibos are labeled according to the following rules. In terms of sentiment polarity, "0" represents positive sentiment, whereas "1"indicates negative sentiment. In terms of the period, "0" indicates the infectious period, while "1" represents the recovered period. The dataset was divided into 4 labeled groups with a total of 12 people involved. Every three members were in one group. Each group was assigned 5000 weibos. Each member in one group was required to label all 5000 weibos without communication to intentionally make the marked category accurate since artificial annotation has a certain subjectivity. The blog was thought to be invalid if the results were inconsistent among the 3 members. After the process of cleaning and arrangement, 15,301 weibos were valid, with the remaining 4699 weibos invalid. We randomly chose 70% of the 15,301 valid weibos as the neural network training set and 30% as the test set.

3. Results

3.1. Relationship between Sentiment Polarity and Flu-Period State at the Word Level

First, a word2vector corpus was built from the 100,000 flu-related weibos, including 50,000 weibos from 2016 and 50,000 weibos from 2017, by removing irrelevant stop words and symbols. The word

frequency of the remaining words was calculated to generate word embedding by similarity and distance between words. Each word consists of a 128-dimensional vector.

One-hundred ninety-five words ranked first and associated with the flu were screened out from the vocabulary and divided into four classes, which represent the two types of flu period (infectious and recovered) and two types of sentiment polarity (positive and negative). The words in the infectious period mainly describe flu confrontation and symptoms. The recovered period describes the status of improvement or remission of the flu. One-hundred ninety-five words in four classes can be found in Table 2.

Table 2. Flu-related Words.

Label	Words
infectious	uncomfortable, not good, ailment, no strength, ache, too awkward, feel bad, serious, sleepy, exhaustion, a tough time, high fever, low fever, pyrexia, diarrhea, emesis, vomit and have watery stools, sneeze, phlegm, dry cough, nasal congestion, difficulty breathing, sore throat, running nose, a bad cold, excessive internal heat, relapse, swell, sore, itch, clinic, children's hospital, emergency call, see a doctor, transfusion, blood, outpatient service, take medicine, injection, drink more water, transfusion, headache, dizzy, backache, weakness in the limbs, stomach ache, leg pain, giddy, terrible, exacerbation, brain swelling, nausea, regurgitation, tonsil, anti-inflammatory drug, capsule, electuary, painful, fester, tinnitus, toothache, sternutation, cough, lacking in strength, intravenous drip, bacteria, influenza, infection, epidemic, the upper respiratory tract, feel chilly, swollen eyes, X-ray, dazed, lethargy, teeter, have a temperature, flu, rhinorrhea, snot, cold cure, inflammation, virus.
recovered	bring down a fever, much better, improve, healthy, recovery, almost gone, feel good, heal, feel all right, get better, fever subsided, antipyretic, abatement of fever, return to normal, stop taking medication, in good health, fitness, get well, improve markedly, pull through, self-cure.
negative	sadness, cry, go crazy, poor, disappointed, tired, heart-broken, agony, worried, unlucky, wronged, grieved, breakdown, disgusting, angry, torturous, sorrow, hard, arduous, piercing pain, sob, anxious, self-accusation, vexation, compunction, fear, gripping, exhausted, weep, worn out, fragile, suffering, helplessness, tantalization, nervous, take offence, guilty, regretful, despairing, whiny, harrowing, depressed, annoying, out of sorts, irritated, listless, bad mood.
positive	quiet, happy, thankful, wish, clear up, alive and kicking, hope, expect, laugh, love, smile, delighted, cheer up, ha ha ha, make an effort, lovely, grinning, felicity, warmth, cheerful, strong, glad, excited, pray, bless, impetrate, look forward, chuckle, satisfied, joyful, active, all the best, smooth going, hang on, have fun, yeah, contented, hug, gentle, safe and sound, benediction, grand time, brave, relieved.

To display the distribution of the 195 words in two-dimensional space, we used t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimension of the word embedding. The scatter plot was drawn by the two-dimensional coordinates of each word in Figure 10. Each point is a word, with a total of 195 points. It is clear that the words that denote positive sentiment and the recovered period are clustered together, and the words in the negative sentiment and infectious period are clustered together closely. In addition, between these 195 words, this paper calculates the similarity between every two words of the flu period and the sentiment polarity based on word2vector. Additionally, a similarity greater than 0.6 was reserved, which is represented by the connections in Figure 10. Therefore, as seen from Figure 10, there are four types of connections, which are divided between the infectious period and negative sentiment, infectious period and positive sentiment, recovered period and negative sentiment, and recovered period and positive sentiment. This paper

also performed statistical analysis on these four types of lines in the scatter plot. When the similarity is greater than 0.6, it is found that there are 53 links between the infectious period and negative sentiment and more than 16 links between the infectious period and positive sentiment. Additionally, there are 81 links between the recovered period and positive sentiment and more than 7 links between the recovered period and negative sentiment. Obviously, the infectious period is much more similar and closer to negative sentiment than to positive sentiment. In contrast, the recovered period is much more similar and closer to positive sentiment than to negative sentiment.

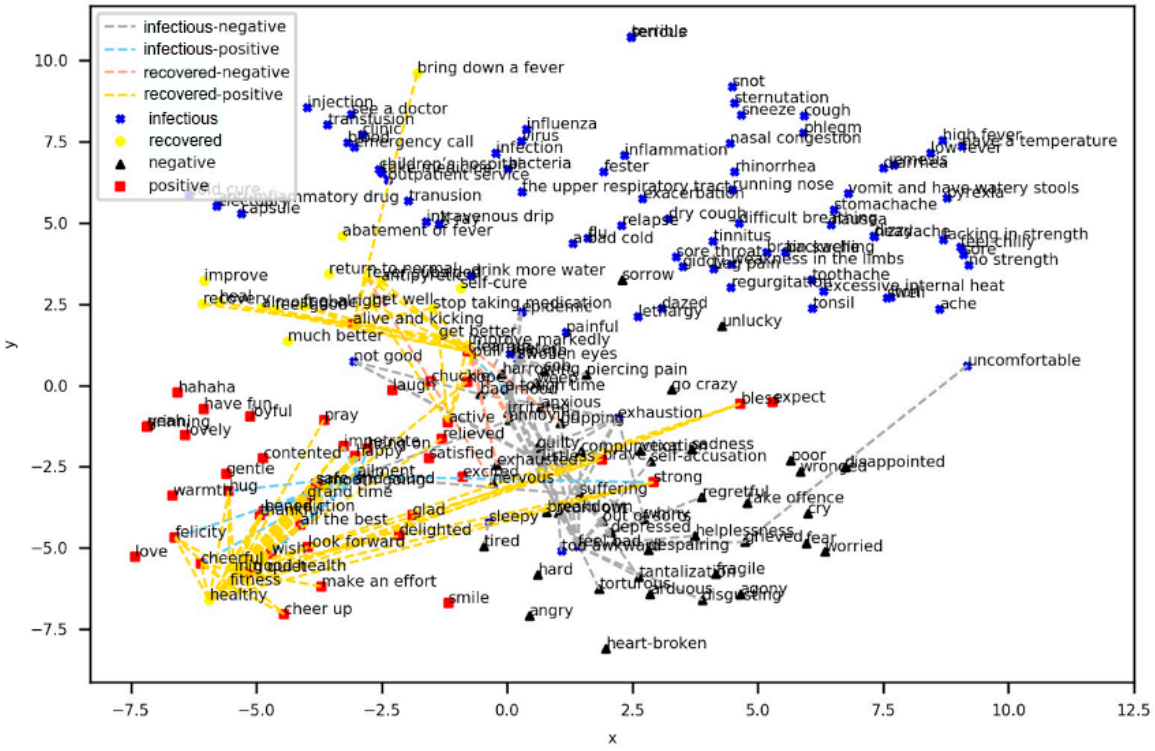


Figure 10. Two-dimensional Word Embedding Scatter Plot.

To demonstrate the relationship between these 195 words and their similarity, a Force Directed Graph of Words' Similarity is shown in Figure 11, where a total of 195 nodes represent the 195 words. The nodes are divided into four categories, and the edges between the nodes are also divided into four categories, the same as in the legend in Figure 10. The force of the nodes is the similarity between the two words. It can be clearly seen from Figure 11 that most of the edges around recovered nodes, such as health and fitness, are all positive nodes, such as cheerful and happy. Additionally, positive nodes, such as active, alive and kicking, are connected with recovered nodes, such as get well and heal. In addition, as can be seen from the similarity forces, these three recovered nodes—in good health, healthy, and fitness—which indicate a healthy status, are connected to more positives nodes than these words are, which indicates that the blogger is recovering but not yet healthy. This finding shows that the healthier the bloggers are, the more positive sentiments they have. Additionally, infectious nodes, such as feel bad and uncomfortable, are mutually connected to negative nodes, such as bad mood and anxious. Among the infectious nodes, from uncomfortable to not good to feel bad, the more serious the disease is, the more the negative nodes are connected.

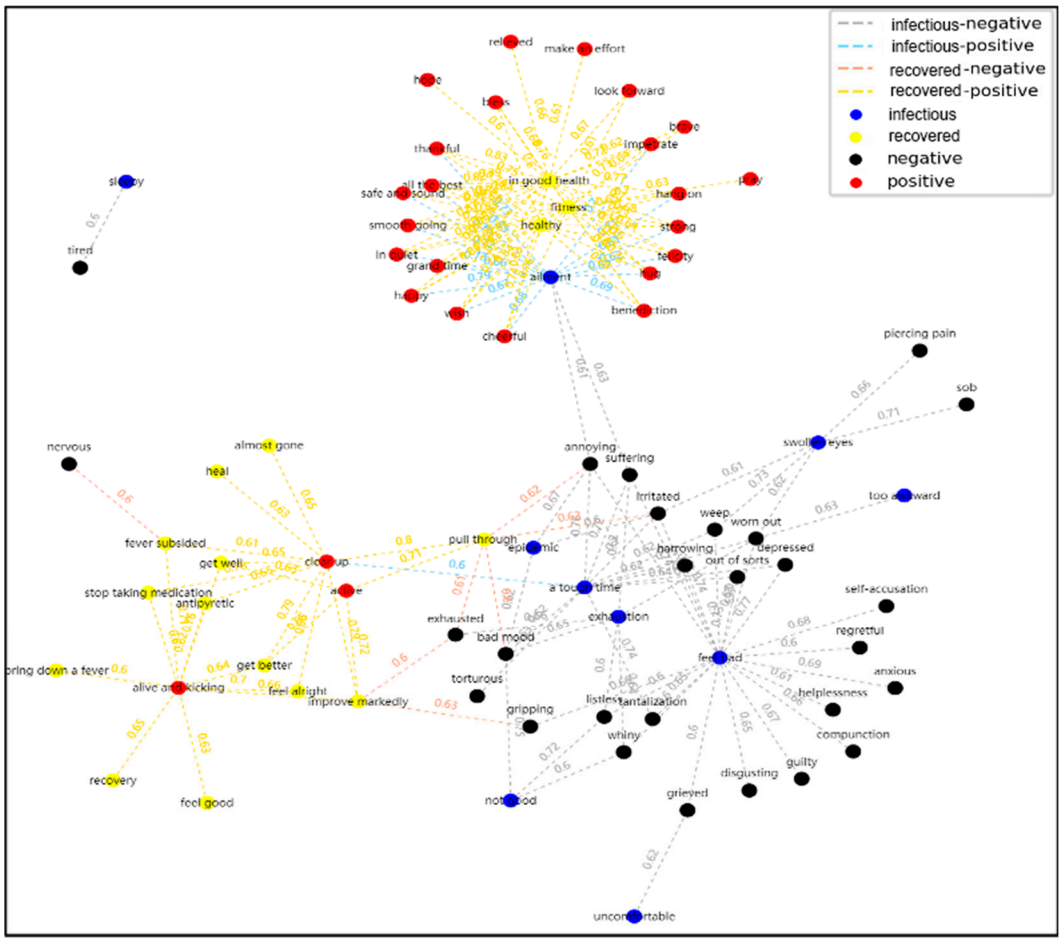


Figure 11. Force-Directed Graph of Words' Similarity.

To locate the word of each point clearly, the scatter of the points for the positive sentiment and recovered period is plotted in Figure 12. The scatter of the points for the negative sentiment and infectious period is shown in Figure 13. It can be seen from Figures 12 and 13 that semantically related words and words with similar meanings are close to each other. Since the number of words that describe the recovered period is significantly lower than the number that describe the infectious period and the meanings of the words that describe the recovered period are mostly similar, it can be determined that the words of the recovered period and positive sentiment are clustered in the second and third quadrants, as shown in Figure 12. The distribution of the words that represent the infectious period and negative sentiment appear to be more scattered in Figure 13, which denotes that these two classes are more relevant.

To measure the interclass distance of these four categories in the scatter, this paper adopts two sample class distances to compare the relationship between the sentiment polarity and the flu period. The first distance is a centroid cluster, which measures the interclass distance by the distance between the two variables' mean value. The coordinates of each class's center gravity are presented in Table 3. The scatter of the class center gravity is plotted through the two-dimensional coordinates of the four categories shown in Figure 14. Figure 14 clearly determines the distribution of four types of class center gravity.

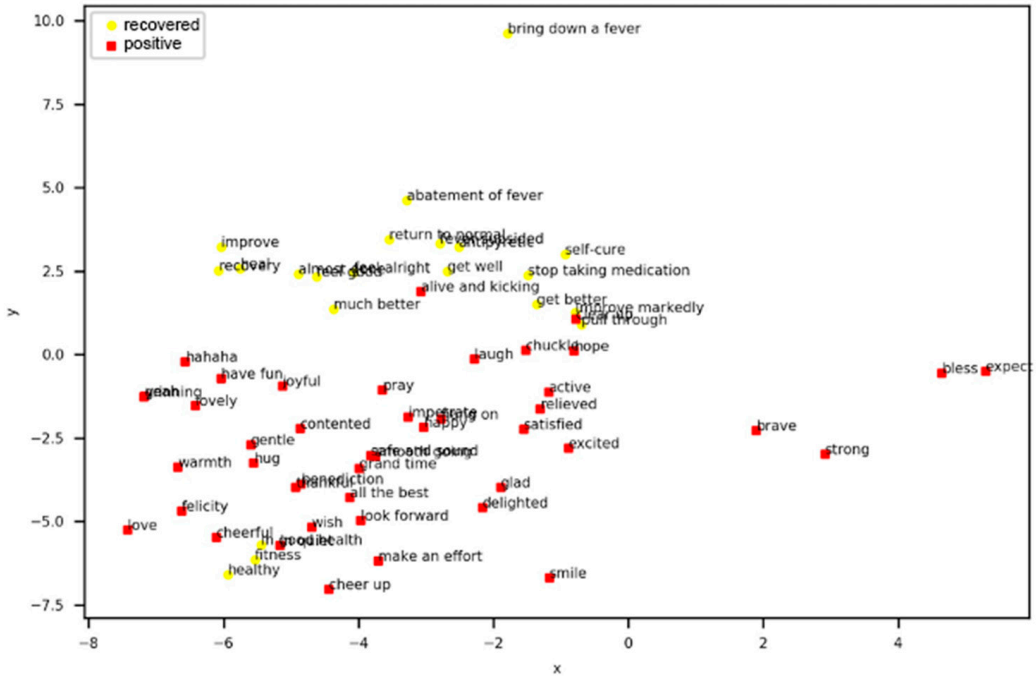


Figure 12. Words of Positive Sentiment and Recovered Period.

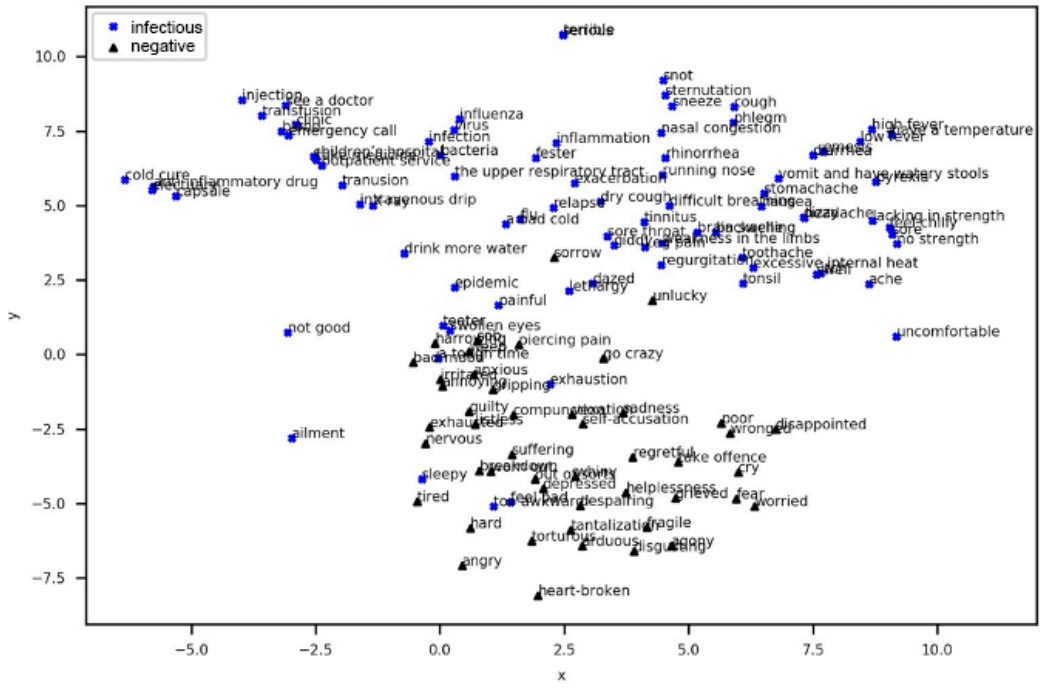


Figure 13. Words of Negative Sentiment and Infectious Period.

Table 3. Class Center Gravity Coordinates.

Label	X	Y
Infectious	2.665	4.758
Negative	2.434	−3.105
Positive	−3.302	−2.661
Recovered	−3.553	1.629

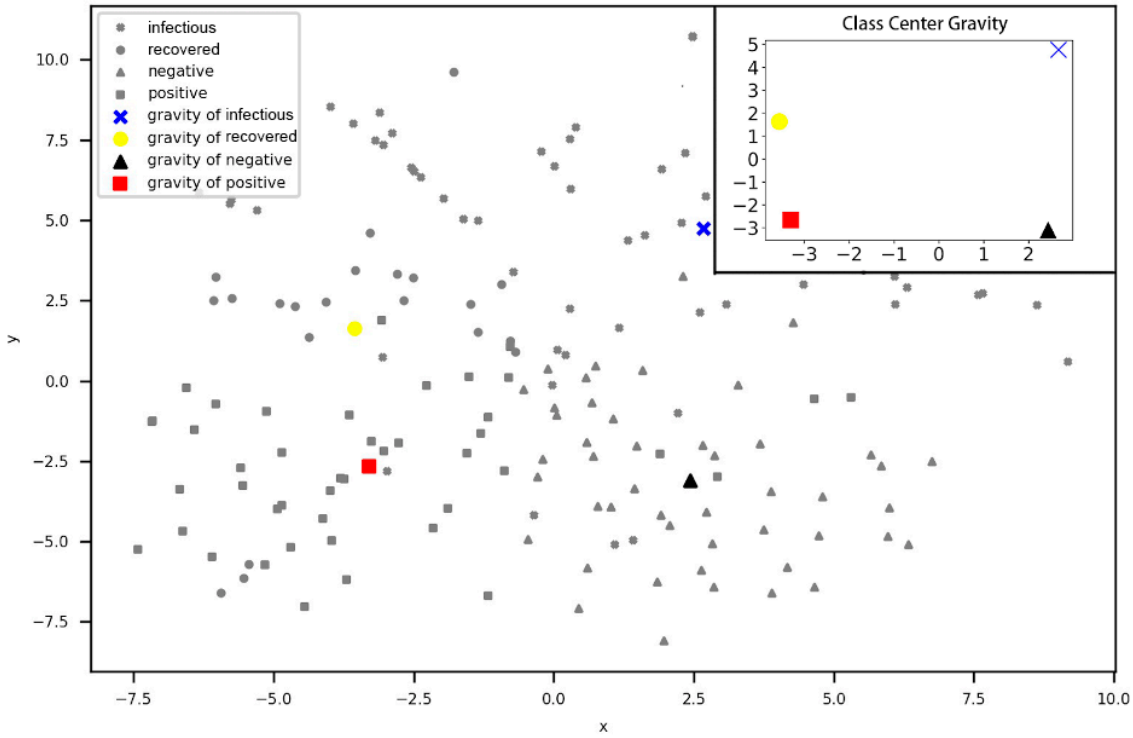


Figure 14. Class Center Gravity Scatter.

Afterwards, we calculated the Euclidean distance matrix of the center gravity in the following section. From the matrix, the distance between the recovered period and positive sentiment is 4.297, while the distance between the recovered period and negative sentiment is 7.632, which significantly shows that the recovered period is closer to positive sentiment. In terms of the infectious period, the distance to the positive sentiment is 9.521, and the distance to the negative sentiment is 7.867, which indicates that the infectious period is closer to the negative sentiment.

$$\begin{array}{c} \\ \text{Recovered} \\ \text{Infectious} \\ \text{Positive} \\ \text{Negative} \end{array} \begin{bmatrix} 0 & 6.960 & 4.297 & 7.632 \\ 6.960 & 0 & 9.521 & 7.867 \\ 4.297 & 9.521 & 0 & 5.753 \\ 7.632 & 7.867 & 5.753 & 0 \end{bmatrix}$$

In addition, another interclass distance measurement method was selected to further verify the correctness of the conclusion, which is known as the between-group linkage and which measures the interclass distance by the average distance between the two categories of individuals. The Euclidean distance matrix is shown in the following section.

$$\begin{array}{c} \\ \text{Recovered} \\ \text{Infectious} \\ \text{Positive} \\ \text{Negative} \end{array} \begin{bmatrix} 0 & 8.739 & 6.364 & 8.724 \\ 8.739 & 0 & 10.806 & 9.376 \\ 6.364 & 10.806 & 0 & 6.977 \\ 8.724 & 9.376 & 6.977 & 0 \end{bmatrix}$$

This finding implies that the conclusion is consistent with the above. The recovered period is closer to the positive sentiment, and the infectious period is closer to the negative sentiment.

3.2. Classification of Flu Period Based on Sentiment Polarity at the Text Level

The flu-related information perceived by social media can detect trend changes and peak points earlier than traditional methods [8,9,19,43]. We also compare the trend of the flu-related weibos ratio

and ILI% from the CDC. It can be seen in Figure 15; the flu data perceived on social media can reflect the trend of official ILI data and report changes and peaks earlier in certain weeks.

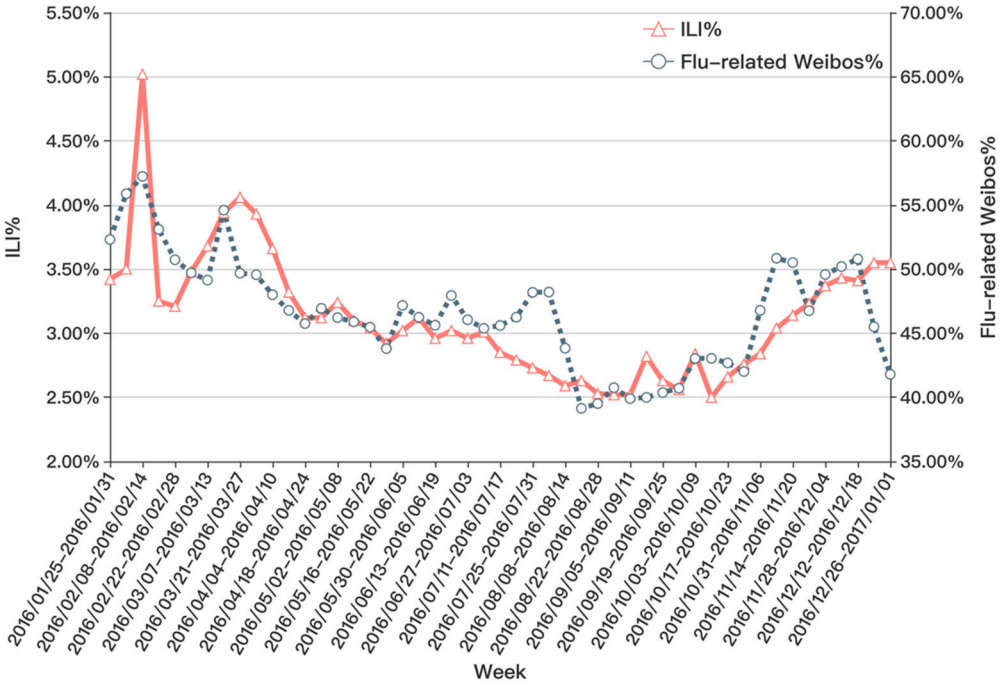


Figure 15. Trend Comparison of Flu-related Weibos% and influenza-like illness (ILI)%.

Two LSTM neural networks were built to classify the sentiment polarity and flu-period in this paper. The input is the word embedding trained by the word2vector based on the corpus of flu-related weibos. It is worthwhile to note that the dimension of the word embedding for the LSTM input is the original word embedding training, which resulted in 128 dimensions. The output of the network is 0 or 1. For the sentiment classification, the output "0" represents negative sentiment, while the output "1" denotes positive sentiment. For the flu-period classification, the output "0" implies that the blogger is infected, while the output "1" implies that the blogger is recovered.

In the LSTM for sentiment classification, the accuracy of the test set reached 0.844 after 30 steps of training. In the LSTM for flu-period classification, the accuracy of the test set reached 0.876 after 30 steps of training. The statistical result of the test set of 4590 weibos was counted to compare the relationship between the flu-period and the sentiment polarity. The prediction result was 876, in which the weibos simultaneously show positive sentiment and the recovered period, shown in Table 3; the other types of prediction results are also shown in Table 3. Apart from these findings, we also counted the number of correct weibos to predict the flu-period in the four types of prediction results. The statistics are shown in Table 4.

Table 4. LSTM Test Set Result.

Period and Sentiment		Positive	Negative	Total
Recovered	total	876	402	1278
	correct	655	176	831
Infectious	total	356	2956	3312
	correct	297	2893	3190

The overall accuracy rate is shown as follows,

$$\mathrm{Acc}_1 = \frac{(831 + 3190)}{(1278 + 3312)} = 0.876 \quad (7)$$

Weibos for both positive sentiment and the recovered period and weibos for both negative sentiment and the infectious period were predicted by the LSTM neural network for a total of 3832 pieces. The accuracy rate is calculated as follows,

$$\mathrm{Acc}_2 = \frac{(655 + 2893)}{(876 + 2956)} = 0.926 \quad (8)$$

Compared to the two results, it can be determined that the accuracy rate increases to 0.926 when the result of the sentiment classification is added, which indicates that the flu period has a certain correlation with sentiment polarity and the classification accuracy of the flu period improves.

4. Discussion

The shortcomings of traditional data are evident since they are manually collected and time-consuming, which leads to high labor costs [8,9]. In addition, traditional methods based on clinical data make it challenging to shed light on the current situation and predict future developing trends [41]. Along with the widespread use of the Internet, social networking data, including web-based epidemiological data, have had explosive growth. Data from social networks show apparent advantages in several respects, such as being real-time and having time-sharing, along with a broad scope of data coverage [31–34]. In terms of the scale of users, Sina Weibo's monthly active users reached 431 million on 30 June 2018, overtaking Twitter, which makes Sina Weibo the world's largest independent social media platform [67]. Sina Weibo is the most popular social media platform in China for the public to share opinions and disseminate information about emergencies and major social events. Therefore, Weibo has a far-reaching scope of dissemination and is an important social influence. Therefore, the use of web-based social media data growth is an imperative trend to use for effective disease control and prevention. Weibo messages carry rich and meaningful implications. Previous approaches in flu state detection through social media have yielded outstanding achievements but have some limitations at the same time. Most obviously, the semantic information was seldom considered; this information might be important for flu detection [50].

However, to our knowledge, this area of study has serious limitations. For example, bloggers experience different flu periods of the latent, infectious, or recovered kind, and their sentiments correspond to the different periods. In reality, most bloggers who are ill are usually negative, while bloggers in recovered periods are active and optimistic. Therefore, omitting the key flu period and non-differentiating Weibo messages data could lead to data contamination and misleading conclusions. It is important to investigate the relationship between the flu period and the sentiment polarity, to make it possible to be conducive to accuracy for classifying the flu period, which directly results in accurately estimating the number of patients in different flu periods. This approach would also help the CDC to take early action for disease control and prevention.

This paper aims to detect the flu period with sentiment polarity at the word and text level based on Sina Weibo data (web-based social media platform), and it proposes optimization suggestions for optimizing the disease detecting process. Several important findings are produced. (1) Social media is a promising and powerful data platform to detect flu patients by earlier discovery rather than traditional medical data. Their periods can be further sorted into infectious and recovered, as mined from social media. (2) The semantic information varies from the weibo texts posted by patients in different flu periods. The interclass distance between the recovered period and positive sentiment is closer than between the recovered period and negative sentiment, and the interclass distance between the infectious and the negative is closer than that of the positive. Additionally, it was noted that the healthier the bloggers are, the more positive sentiments they have. The more serious the flu is, the more that bloggers are connected with negative emotions. (3) A multichannel disease detection model is developed in this study to evaluate and classify the flu period with an accuracy of up to 0.926 based on the LSTM network. Our optimized model effectively improves the classification accuracy of the flu period after adding the sentiment classification results.

The research findings have important theoretical implications. (1) The previous literature investigates the sentiment and disease predictions separately. This paper examines the relationship between sentiment and disease detection. We found that by adding sentiment factors, the classification accuracy is improved remarkably, from 0.876 to 0.926. (2) This paper explores the relationship between the sentiment polarity and the flu-period at two levels of words and text, combining the methods of word2vector and LSTM, which have been used rarely for disease surveillance studies. (3) This research proposes a complete theoretical framework based on web-based social media data. The use of this model can be extended to many aspects, such as monitoring chronic and mental diseases.

This study also has important practical implications. (1) This paper optimizes the disease detecting process and establishes multichannel surveillance measures for CDC decision making. (2) This paper will monitor a larger range of infected population. Furthermore, it can identify patients in advance who are not aware of disease. (3) The previous weibo text processing classifies only flu-related weibos and unrelated weibos. This paper further divides flu-related weibos into two periods: recovered and infectious. Research outcomes improve the reliability and accuracy for the prediction of flu trends. Both point (2) and point (3) not only help the CDC to detect disease information in real time but also provide a novel method for disease information management. (4) The conclusion supports the expansion of the number of neural network training sets, eliminating some of the high cost of manual labeling. The classification results of the flu-period can be replaced in the model to increase the amount of training set data, which enables the LSTM neural network to fully learn to better characterize the model.

5. Conclusions

Timely and reliable flu monitoring is an important basis for successful control of the spread of disease and mitigation of the associated damage. However, due to its high contagiousness and rapid spread, the flu epidemic has caused great difficulties in prevention and surveillance. With the rapid development and popularity of web-based social media platform data, Sina Weibo, one of the world's largest social media companies, has become an ideal data source to make real-time, low-cost surveillance possible as an early warning of outbreaks and an adjunct to traditional methods of investigation. According to the latest estimate by the United States Centers for Disease Control and Prevention (US-CDC), as many as 650,000 people worldwide die from seasonal flu-related respiratory diseases each year. It is evident that the flu imposes a heavy burden on the international community, and the flu's global, social and economic costs are considerable. It is worth noting that improving the ability to monitor infectious disease is the key to further strengthening management capacity of the health system and organizing a massive flu outbreak response.

However, most traditional epidemiological surveillance methods adopt clinical data through manual information collection, showing the shortcomings of high labor costs but also causing a lag in data timeliness. The data limitation makes it difficult to understand the current situation, which is critical for flu trend forecasting. Social media data increase at a rapid pace, including epidemiological data, which offer benefits in terms of timeliness and magnitude. Sina's 2018 quarter earnings report that the number of monthly active users of Sina Weibo is larger than that of Twitter, which makes it the world's largest independent social media company in terms of user scale. Sina Weibo has always been the starting data source of various types of major emergencies in China, and its commercial value has been continuously promoted and has great potential.

This paper explores the relationship between the flu-period and sentiment polarity from two levels based on Sina Weibo data. To be specific, at the word level, we used word2vector to create the flu-related weibos corpus and the t-SNE method to reduce the dimension. The centroid cluster and between-group linkage were jointly used to measure the distance between the four classes, thus visually showing the relationship between the sentiment polarity and flu-period. At the text level, the sentiment polarity and flu-period of flu-related weibos were classified by the LSTM networks, respectively. We counted the classification results as both belonging to the infectious and negative sentiment as well

as to the recovered and positive sentiment, and we calculated the accuracy rate. We then compared the rate with the overall flu-period classification accuracy to observe the differences. This paper proposes an integrated conceptual framework and practical methods for optimizing the disease detection process with fast information, early discovery, added infected cases and high accuracy. These contributions are described in detail as follows:

First, in theory, this paper integrates various channels for detecting infectious diseases in real time with fast information. In addition to the clinical data and search engine data, the detecting data obtained through social media can also provide prompt and time-sharing disease information to the Centers for Disease Control and Prevention (CDC). The monitoring mechanisms operate in real time, which can help the CDC fully prepare for the next round of prevention and control.

Second, in practice, social media enables the early discovery of disease infection. The sooner the disease is diagnosed, the easier it is to properly treat and controlled. The CDC is committed to pursuing early detection of diseases. Through social media platforms, we can detect the spread and severity of a disease earlier than search engines and the CDC. When diseases break out, the patient might not be aware of them but could post on Twitter or Weibo. The behavior would be recorded by social media sensors. Based on human behavioral theory, the data possess unique value for detecting disease trends.

Third, social media is adept at tracking more patients than traditional clinic data. Larger infectious populations can be monitored by social media than with clinic data. Influenza-like illnesses (ILI) published by the CDC are measured according to outpatient statistics when fevers are higher than 38 degrees and are accompanied by a cough or sore throat. However, a considerable number of people often choose not to go to the hospital for treatment when they have the flu or might buy medicine from a pharmacy by themselves, which cannot be counted in ILI measurements. Social media can detect these patients, which could result in a larger amount of meaningful data being collected, and thus, these data could lead to more reliable prediction of disease outbreaks.

Fourth, this paper detects disease periods with observably high accuracy, which could directly result in significant differences in treatment and disease control measures. Targeting the disease period precisely helps clinical managers to improve the treatment effect and reduces the prevention cost by rationally allocating resources, such as medical personnel and medicine as well. This paper can not only detect whether the patient has the flu but also classify the flu period, infectious or recovered period, which lays the foundation for predicting future flu trends. It also provides another data source to assist the CDC in managing disease information.

Fifth, in terms of theoretical contributions, this paper investigates the relationship between sentiment polarity and the flu period at different word and text levels by combining the word2vector and LSTM methods, thereby carrying out interdisciplinary research in the fields of sentiment analytics and health informatics. In addition, this paper provides an effective solution for artificially labeling a training set. High-accuracy weibo texts can be used to boost the size of the training set, thus saving time and labor costs.

In future work, we need to study a wider range of data since the current data only cover two years, 2016 and 2017. Moreover, this paper compares the trend of official ILI data from the CDC and flu-related data from social media in 2016. We will examine more valuable disease information from social media-based data on a larger scale.

Author Contributions: Conceptualization, S.S. and Q.Y.; methodology, S.S. and Q.Y.; software, S.S.; validation, S.S. and Y.W.; formal analysis, Q.Y.; investigation, Q.Y.; resources, S.S.; data curation, S.S.; writing—original draft preparation, Q.Y.; writing—review and editing, S.S. and Y.W.; visualization, Q.Y.; supervision, S.S.; project administration, S.S.; funding acquisition, S.S. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant numbers are 71771010, 72071010 and 71904009.

Conflicts of Interest: 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.

References

1. Sidana, S.; Amer-Yahia, S.; Clausel, M.; Rebai, M.; Mai, S.T.; Amini, M.R. Health monitoring on social media over time. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1467–1480. [CrossRef]
2. Sinnenberg, L.; DiSilvestro, C.L.; Mancheno, C.; Dailey, K.; Tufts, C.; Buttenheim, A.M.; Barg, F.; Ungar, L.; Schwartz, H.; Brown, D.; et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiol.* **2016**, *1*, 1032–1036. [CrossRef]
3. Centers for Disease Control and Prevention of the United States of America. Available online: https://www.cdc.gov/flu/weekly/pastreports.htm (accessed on 21 August 2018).
4. Belser, J.A.; Tumpey, T.M. The 1918 flu, 100 years later. *Science* **2018**, *359*, 255. [CrossRef]
5. Hasnain, S.E. Molecular epidemiology of infectious diseases: A case for increased surveillance. *Bull. World Health Organ.* **2003**, *81*, 474. [CrossRef]
6. Summary table of SARS cases by country, 1 November 2002–7 August 2003. Available online: http://www.who.int/csr/sars/country/2003_08_15/en/ (accessed on 15 August 2003).
7. Smith, R.D. Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Soc. Sci. Med.* **2006**, *63*, 3113–3123. [CrossRef]
8. Wang, F.; Wang, H.; Xu, K.; Raymond, R.; Chon, J.; Fuller, S.; Debruyn, A. Regional Level Influenza Study with Geo-Tagged Twitter Data. *J. Med. Syst.* **2016**, *40*, 1–8. [CrossRef]
9. Allen, C.; Tsou, M.-H.; Aslam, A.; Nagel, A.; Gawron, J.-M. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PLoS ONE* **2016**, *11*, e0157734. [CrossRef] [PubMed]
10. Yan, L. Good Intentions, Bad Outcomes: The Effects of Mismatches between Social Support and Health Outcomes in an Online Weight Loss Community. *Prod. Oper. Manag.* **2018**, *27*, 9–27. [CrossRef]
11. Rolls, K.; Hansen, M.; Jackson, D.; Elliott, D. How health care professionals use social media to create virtual communities: An integrative review. *J. Med. Internet Res.* **2016**, *18*, e166. [CrossRef] [PubMed]
12. Lau, R.Y.K.; Zhang, W.; Xu, W. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Prod. Oper. Manag.* **2018**, *27*, 1775–1794. [CrossRef]
13. Wu, D.; Cui, Y. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decis. Support Syst.* **2018**, *111*, 48–59. [CrossRef]
14. Wang, Z.; Ye, X. Social media analytics for natural disaster management. *Int. J. Geogr. Inf. Sci.* **2017**, *2*, 1–24. [CrossRef]
15. Shan, S.; Liu, X.; Wei, Y.; Xu, L.; Zhang, B.; Yu, L. A new emergency management dynamic value assessment model based on social media data: A multiphase decision-making perspective. *Enterp. Inf. Syst.* **2020**, *14*, 680–709. [CrossRef]
16. Hamilton, L.A.; Franks, A.; Heidel, R.E.; McDonough, S.L.K.; Suda, K.J. Assessing the Value of Online Learning and Social Media in Pharmacy Education. *Am. J. Pharm. Educ.* **2016**, *80*, 97. [CrossRef] [PubMed]
17. Butler, D. When google got flu wrong. *Nature* **2013**, *494*, 155–156. [CrossRef]
18. Lazer, D.; Kennedy, R.; King, G.; Vespignani, A. The parable of google flu: Traps in big data analysis. *Science* **2014**, *343*, 1203–1205. [CrossRef]
19. Update: Influenza Activity in the United States During the 2017–18 Season and Composition of the 2018–19 Influenza Vaccine. Available online: https://www.cdc.gov/mmwr/volumes/67/wr/mm6722a4.htm?s_cid=mm6722a4_w (accessed on 8 June 2018).
20. Yun, G.W.; Morin, D.; Park, S.; Joa, C.Y.; Labbe, B.; Lim, J.; Lee, S.; Hyun, D. Social media and flu: Media Twitter accounts as agenda setters. *Int. J. Med. Inform.* **2016**, *91*, 67–73. [CrossRef]
21. Saldana-Perez, A.M.M.; Moreno-Ibarra, M. Traffic analysis based on short texts from social media. *Int. J. Knowl. Soc. Res.* **2016**, *7*, 63–79. [CrossRef]
22. Liang, S.; Ren, Z.; Zhao, Y.; Ma, J.; Yilmaz, E.; De Rijke, M. Inferring Dynamic User Interests in Streams of Short Texts for User Clustering. *Acm Trans. Inf. Syst.* **2017**, *36*, 1–37. [CrossRef]
23. Tommasel, A.; Godoy, D. A Social-aware online short-text feature selection technique for social media. *Inf. Fusion* **2018**, *40*, 1–17. [CrossRef]

24. Heesterbeek, H.; Anderson, R.M.; Andreasen, V.; Bansal, S.; De Angelis, D.; Dye, C.; Eames, K.T.D.; Edmunds, W.J.; Frost, S.D.W.; Funk, S.; et al. Modeling infectious disease dynamics in the complex landscape of global health. *Science* **2015**, *347*, aaa4339. [CrossRef] [PubMed]
25. Rus, H.M.; Cameron, L.D. Health Communication in Social Media: Message Features Predicting User Engagement on Diabetes-Related Facebook Pages. *Ann. Behav. Med.* **2016**, *50*, 678–689. [CrossRef] [PubMed]
26. Long, E.F.; Nohdurft, E.; Spinler, S. Spatial Resource Allocation for Emerging Epidemics: A Comparison of Greedy, Myopic, and Dynamic Policies. *Manuf. Serv. Oper. Manag.* **2018**, *20*, 181–198. [CrossRef]
27. Chen, Q.; Ayer, T.; Chhatwal, J. Optimal M-Switch Surveillance Policies for Liver Cancer in a Hepatitis C–Infected Population. *Oper. Res.* **2018**, *66*, 673–696. [CrossRef]
28. Ozaltin, O.Y.; Prokopyev, O.A.; Schaefer, A.J. Optimal Design of the Seasonal Influenza Vaccine with Manufacturing Autonomy. *Inf. J. Comput.* **2018**, *30*, 371–387. [CrossRef]
29. Duijzer, L.E.; van Jaarsveld, W.L.; Wallinga, J.; Dekker, R. Dose-Optimal Vaccine Allocation over Multiple Populations. *Prod. Oper. Manag.* **2018**, *27*, 143–159. [CrossRef]
30. Lee, E.C.; Arab, A.; Goldlust, S.M.; Viboud, C.; Grenfell, B.T.; Bansal, S. Deploying digital health data to optimize influenza surveillance at national and local scales. *PLoS Comput. Biol.* **2018**, *14*, e1006020. [CrossRef]
31. Tambo, E.; Adetunde, O.T.; Olalubi, O.A. Re-emerging lassa fever outbreaks in Nigeria: Re-enforcing "one health" community surveillance and emergency response practice. *Infect. Dis. Poverty* **2018**, *7*, 37. [CrossRef]
32. Ruomeng, C.; Santiago, G.; Antonio, M.; Zhang, D.J. The operational value of social media information. *Prod. Oper. Manag.* **2017**, *27*, 1749–1769. [CrossRef]
33. Pandey, A.C.; Rajpoot, D.S.; Saraswat, M. Twitter sentiment analysis using hybrid cuckoo search method. *Inf. Process. Manag.* **2017**, *53*, 764–779. [CrossRef]
34. D'Andrea, E.; Ducange, P.; Lazzerini, B.; Marcelloni, F. Real-time detection of traffic from twitter stream analysis. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2269–2283. [CrossRef]
35. Aiello, A.E.; Renson, A.; Zivich, P.N. Social Media- and Internet-Based Disease Surveillance for Public Health. *Annu. Rev. Public Health* **2020**, *41*, 101–118. [CrossRef] [PubMed]
36. Shan, S.; Zhao, F.; Wei, Y.; Liu, M. 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. [CrossRef]
37. Raamkumar, A.S.; Tan, S.G.; Wee, H.L. Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook during the COVID-19 Pandemic in Early 2020: Cross-Country Comparison. *J. Med. Internet Res.* **2020**, *22*, 12. [CrossRef]
38. Lwin, M.O.; Lu, J.H.; Sheldenkar, A.; Schulz, P.J. Strategic Uses of Facebook in Zika Outbreak Communication: Implications for the Crisis and Emergency Risk Communication Model. *Int. J. Environ. Res. Public Health* **2018**, *15*, 19. [CrossRef]
39. Vijaykumar, S.; Meurzec, R.W.; Jayasundar, K.; Pagliari, C.; Fernandopulle, Y. What's buzzing on your feed? Health authorities' use of Facebook to combat Zika in Singapore. *J. Am. Med. Inf. Assoc.* **2017**, *24*, 1155–1159. [CrossRef]
40. Dubey, D.; Amritphale, A.; Sawhney, A.; Dubey, D.; Srivastav, N. Analysis of YouTube as a source of information for West Nile Virus infection. *Clin. Med. Res.* **2014**, *12*, 129–132. [CrossRef]
41. Davidson, M.W.; Haim, D.A.; Radin, J.M. Using Networks to Combine "Big Data" and Traditional Surveillance to Improve Influenza Predictions. *Sci. Rep.* **2015**, *5*, 8154. [CrossRef]
42. Chen, L.Z.; Hossain, K.; Butler, P.; Ramakrishnan, N.; Prakash, B.A. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Min. Knowl. Discov.* **2016**, *30*, 681–710. [CrossRef]
43. Lamb, A.; Paul, M.J.; Dredze, M. Separating fact from fear: Tracking flu infections on twitter. In Proceedings of the NAACL, Atlanta, Georgia, 9–14 June 2013; pp. 789–795.
44. Wang, P.; Xu, B.; Xu, J.; Tian, G.; Liu, C.; Hao, H. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing* **2016**, *174*, 806–814. [CrossRef]
45. Muhammad, A.; Wiratunga, N.; Lothian, R. Contextual sentiment analysis for social media genres. *Knowl. -Based Syst.* **2016**, *108*, 92–101. [CrossRef]
46. Scarpa, G.; Gargiulo, M.; Mazza, A.; Gaetano, R. A CNN-Based Fusion Method for Feature Extraction from Sentinel Data. *Remote Sens.* **2018**, *10*, 236. [CrossRef]

47. Jiang, D.; Luo, X.; Xuan, J.; Xu, Z. Sentiment Computing for the News Event Based on the Social Media Big Data. *IEEE Access* **2017**, *5*, 2373–2382. [CrossRef]
48. Shan, S.; Peng, J.; Wei, Y. Environmental Sustainability assessment 2.0: The value of social media data for determining the emotional responses of people to river pollution—A case study of Weibo (Chinese Twitter). *Socio-Econ. Plan. Sci.* **2020**. [CrossRef]
49. Chen, Z.; Zhang, R.; Xu, T.; Yang, Y.; Wang, J.; Feng, T. Emotional attitudes towards procrastination in people: A large-scale sentiment-focused crawling analysis. *Comput. Hum. Behav.* **2020**, *110*, 106391. [CrossRef]
50. Sun, X.; Ye, J.; Ren, F. Detecting influenza states based on hybrid model with personal emotional factors from social networks. *Neurocomputing* **2016**, *210*, 257–268. [CrossRef]
51. Adamopoulos, P.; Ghose, A.; Todri, V. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. *Inf. Syst. Res.* **2018**, *29*, 612–640. [CrossRef]
52. Lee, D.; Hosanagar, K.; Nair, H.S. Advertising content and consumer engagement on social media: Evidence from Facebook. *Manag. Sci.* **2018**, *64*, 5105–5513. [CrossRef]
53. Li, T.; Mei, T.; Kweon, I.-S.; Hua, X.-S. Contextual Bag-of-Words for Visual Categorization. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 381–392. [CrossRef]
54. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
55. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537. [CrossRef]
56. Cui, P.; Wang, X.; Pei, J.; Zhu, W. A survey on network embedding. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 833–852. [CrossRef]
57. Heimbach, I.; Hinz, O. The Impact of Sharing Mechanism Design on Content Sharing in Online Social Networks. *Inf. Syst. Res.* **2018**, *29*, 592–611. [CrossRef]
58. Timoshenko, A.; Hauser, J.R. Identifying customer needs from user-generated content. *Mark. Sci.* **2019**, *38*. [CrossRef]
59. Hughes, M.; Li, I.; Kotoulas, S.; Suzumura, T. Medical text classification using convolutional neural networks. *Stud. Health Technol. Inform.* **2017**, *235*, 246–250. [CrossRef] [PubMed]
60. Kadetotad, D.; Yin, S.; Berisha, V.; Chakrabarti, C.; Seo, J.-S. An 8.93 TOPS/W LSTM Recurrent Neural Network Accelerator Featuring Hierarchical Coarse-Grain Sparsity for On-Device Speech Recognition. *IEEE J. Solid-State Circuits* **2020**, *55*, 1877–1887. [CrossRef]
61. Hochreiter, S.; Schmidhuber, J. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 1997, pp. 1735–1780. [CrossRef]
62. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. Sentiment analysis of Twitter data. The Workshop on Languages in Social Media. *Assoc. Comput. Linguist.* **2011**, *39*, 30–38. [CrossRef]
63. Taboada, M.; Brooke, J.; Tofiloski, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
64. Gao, W.; Su, C. Analysis on block chain financial transaction under artificial neural network of deep learning. *J. Comput. Appl. Math.* **2020**, *380*, 112991. [CrossRef]
65. Gers, F.A.; Schraudolph, N.N. Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **2003**, *3*, 115–143. [CrossRef]
66. The 41th China Statistical Report on Internet Development. Available online: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249.htm (accessed on 5 March 2018).
67. Weibo Posts Unaudited Earnings for the Second Quarter in 2018. Available online: https://tech.sina.com.cn/i/2018-08-08/doc-ihhkuskt9159883.shtml (accessed on 8 August 2018).