Chinese Social Media Analysis for Disease Surveillance

Nanhai Yang[1], Xiaohui Cui[1]*, Cheng Hu[1], Weiping Zhu[1], Chengrui Yang[1]
[1]International School of Software, Wuhan University
Wuhan, Hubei, China
Email: xcui@whu.edu.cn

***Abstract*—It is reported that there are hundreds of thousands of deaths caused by seasonal flu all around the world every year. More other diseases such as chickenpox, malaria, etc. are also serious threat to people's physical and mental health. Therefore proper techniques for disease surveillance are highly demanded. Recently, social media analysis is regarded as an efficient way to achieve this goal, which is feasible since growing number of people post their health information to social media such as blogs, personal website, etc. Previous work on social media analysis mainly focused on English materials but hardly considered Chinese materials, which hinders the use of such technique for Chinese people. In this paper, we proposed a new method of Chinese social media analysis for disease surveillance. More specifically, we compared different kinds of methods in the process of classification, and then proposed a new way to process Chinese text data. The Chinese Sina micro-blog data collected from September to December 2013 is used to validate the effectiveness of the proposed method. The results show that a high classification precision of 87.49% in average is obtained. Comparing with the data from the authority, Chinese national influenza center, we can predict the outbreak time of flu 5 days earlier.**

Keywords—Social media; Chinese, SVMLIGHT; classification; prediction; flu

I. INTRODUCTION

With the popularity and development of Internet, new social media such as blogs, personal website, instant messaging, etc. has been greatly changing people's life. These social media propelling the spread of social news, public opinion and personal daily information in human's society, playing an important role in current information dissemination. According to the survey of iResearch (as shown in figure 1), the time that people spend on social media reaches 4.6 hours every week by the end of September 2009[2]., and it will keep increasing in the future[3].

Since people spend so much time on social media, it is worthy to utilize them to collect and uncover various kinds of health information. There are many researchers engaging in the analysis of English social media for disease surveillance or related works. In 2006, a website called "who is sick" is developed for people to post their sickness information [4]. In 2008, Jeremy Ginsber predicted flu 1 to 2 week earlier than Centers for Disease Control (CDC) through analyzing the log files of google search [5]. After that, more researchers predicted the outbreak time of diseases by using internet data [6-8]. Ficeifeld C C collects users' health information through an application installed on mobile phones and then detects diseases [9]. Adam Sadilek analyzes the content of Twitter of the users and their friends, and then use them to predict the users' body health status [1011]. However, few researches have been done by analyzing Chinese social media. . Zhengyan Cui classified short text into different kinds of categories (sports, news and etc.)using K Nearest Neighbor (KNN)[12]. Yang F proposed a method of automatic detection of rumor Sina micro-blogs [13]. These works are not for disease surveillance. To the best of our knowledge, there are no researches on disease surveillance through mining Chinese social media.

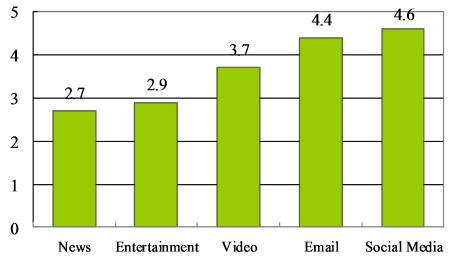

Fig 1. Global Internet users' online time assignments in 2010

This paper aims to predicate people's health status in a region of Beijing based on Sina micro-blog, a famous China social medium, . We collected the related data and propose an effective classification method for such purpose. As a result, we can classify the micro-blog data with 87.49% precision and 92.28% recall in average.

This work is an important step towards predicting disease based on Chinese social media. It explores the process of classification of short text, and compares different kinds of method in the process of classification to get a better result. It also provides foundation for researches on predicting disease by analyzing social media information.

The rest of the paper is organized as follows: Section II presents the procedure of this research. Section III describes how to get the data and the character of this data.

978-1-4799-8003-1/14 $31.00 © 2014 IEEE
DOI 10.1109/IIKI.2014.11

Section IV discusses our simulation environment and Section V describes the method used in this research procedure and the experiment result. Finally, section V concludes the paper.

II. RESEARCH PROCEDURE

Our research flow is shown in Figure 2. We begin by feasibility analysis, and found no researches have been done in disease surveillance on Chinese Social Media to the best of our knowledge; then we start to collecting the data from Sina and have the data cleaned; after that we select sample data randomly and classify them manually; we build upon previous work on classification of text messages (k-means, KNN, SVM), to get a better result with a relatively better classifier, and apply those methods into our classification.

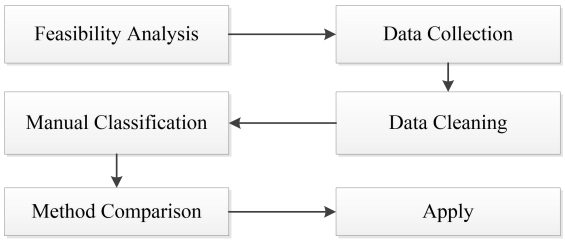

Fig 2. Procedures of research

III. THE DATA

Micro-blog is a kind of blog service through which people can post their messages no more than 140 characters. It makes users to express their thought briefly and encourages frequent information updates. In China, there are mainly three such kind of micro-blogs: Sina micro-blog (http://weibo.com/), Tencent micro-blog (http://t.qq.com/), NetEase micro-blog (http://t.163.com/) and sohu micro-blog (http://t.sohu.com/). According to the analysis of google trend from 2010(Figure 3), Sina micro-blog is the most popular Chinese social media, and our analysis and evaluation are based on the data obtained from Sina micro-blog.

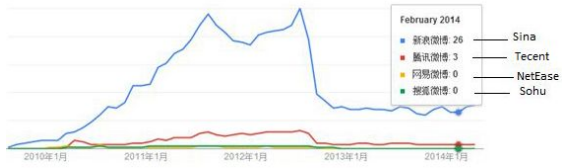

Fig 3. A contrasts of interest in Chinese micro-blog

Sina provides convenient API for obtaining data, such 'Public Micro-blog'; 'Location Based Micro-blog'; 'Location Nearby Micro-blog'; etc. 'Public Micro-blog' provides all the micro-blogs without location information; 'Location Based Micro-blog' provides geotag micro-blog in specific place, it cannot covers most of the micro-blog; 'Location Nearby Micro-blogs' provides geotag micro-blog, and can cover the micro-blog over the area we choose.

We choose 'location nearby micro-blog' (Its main parameters for the APIs are listed in Table 1) because we want to get all the micro-blogs in a specific area with location information [14]. We choose a circle area in Beijing as shown in Figure 4.

Using 'location nearby micro-blog' API and Java scripts, we collect Sina micro-blogs of that circle area (longitude: 116.39750833333, latitude: 39.908647222222, Range 11120) from September 2013 to December 2013. There are 3505110 pieces of micro-blog in total, which includes 951299 pieces of micro-blog in September, 900337 pieces of micro-blog in October, 861590 pieces of micro-blog in November, and 791884 pieces of micro-blog in December. Our dataset is related to 374411 persons, and and there are 4.4 pieces of micro-blogs per person in average.

Table 1. Parameters of 'location nearby micro-blog'

Parameter	Meaning
Lat	latitude
Long	longitude
Range	Radius of Search range
StartTime	Time start to obtain data, expressed as UNIX timestamp
EndTime	Time end to obtain data, expressed as UNIX timestamp

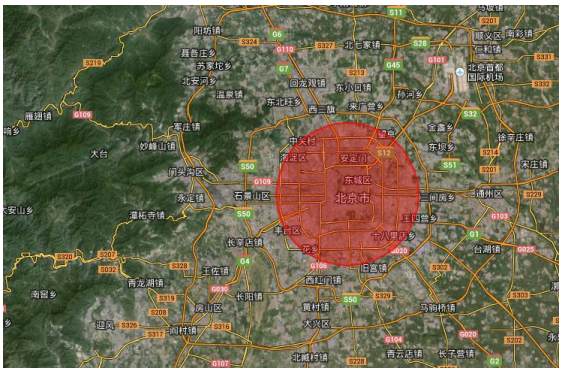

Fig 4. Places and scopes to obtain data

We select 5000 pieces of status randomly and classify them manually into two categories; one is 'sick micro-blog' which indicates the author is sick; the other one is 'not sick micro-blog' which indicates the author is not sick. Among these micro-blogs, we get 285 'sick micro-blog', and select 285 'not sick micro-blog' as training data and test data.

IV. APPROACH

In this section, we follow the steps illustrated in Figure 5 to conduct the prediction. First, we conduct some preprocessing to eliminate the noise information, and then use text model to express the content of every piece of micro-blog. Then, we train the classifier using training data and the results are record into model file. After that, we predicate the flu based on the training model and the test data. Finally, we evaluate our proposed method and report the results.

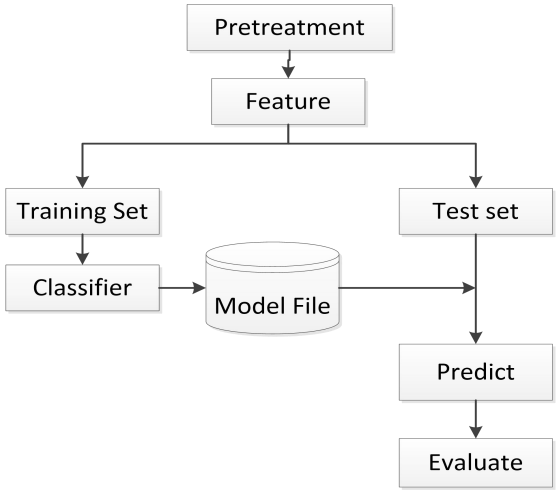

Fig 5. Classification procedures

A. Textual Feature

One important different of text processing between English and Chinese is the extraction of textual Feature. Although there are many Chinese word segmentation systems which can provide word segmentation with a high accuracy, it is not easy to determine whether using word or character as textual features.

This paper uses the same dataset and the same classifier to compare these two kinds of features. According to our result shown in Fig 6 and 7, using word as features can get a higher precision and recall, which reach 0.90 precision and 0.947 recall compared with 0.785 precision and 0.895 recall by using character as features. And we use word as features, for example: a micro-blog 'I am sick' is represented by following feature vector:

(我感冒了)=(我,感冒,了)

Means: (I am sick) = (I, am, sick)

```
Reading model...OK. (366 support vectors read)
Classifying test examples..100..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 82.46% (94 correct, 20 incorrect, 114 total)
Precision/recall on test set: 78.46%/89.47%
```

Fig 6. Result of the classification using character as textual feature

```
Reading model...OK. (410 support vectors read)
Classifying test examples..100..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 92.11% (105 correct, 9 incorrect, 114 total)
Precision/recall on test set: 90.00%/94.74%
```

Fig 7. Result of the classification using word as textual feature

B. Word weighting

Since word is the advanced language used by human, human can gain the information of the text, but computers cannot, and it's necessary to convert the text message into the message computers can understand, and we need to pre compute the weight of different word. Word weighting method is another problem which should be considered when we start to classify micro-blogs. There are four kinds of word weighting method: Boolean weighting, term frequency weighting (TF), inverted document frequency weighting (IDF) and term frequency-inverted document frequency weighting (TFIDF). Boolean weighting does not consider the importance of each word, and term frequency weighting does not consider the entire corpus, and this paper compare IDF with TDIDF, the accuracy of TDIDF (Figure 9) is 90.00% which is much higher than IDF (accuracy=?)(Figure 8), and this paper uses TFIDF as a method of word weighting. [15-16]

```
Reading model...OK. (267 support vectors read)
Classifying test examples..100..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 64.91% (74 correct, 40 incorrect, 114 total)
Precision/recall on test set: 58.76%/100.00%
```

Fig 8. Result of the classification using IDF as word weighting

```
Reading model...OK. (410 support vectors read)
Classifying test examples..100..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 92.11% (105 correct, 9 incorrect, 114 total)
Precision/recall on test set: 90.00%/94.74%
```

Fig 9. Result of the classification using TFIDF as word weighting

C. Classifier

For the classification, SVM and K-means are two methods widely used in the machine learning area. In this paper, we tried both these two approaches and compared their performance for our problem. We get the accuracy of 47.02% for SVM and 90.00% for K-means. We think the reason for that is we can provide specific labels for K-means in this problem which facilitates the classification

and has a better performance. Therefore we adopt K-means for our classification.

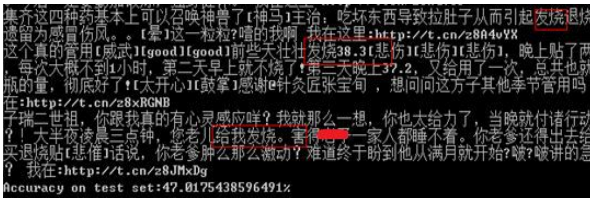

Fig 10. Result of K-means classification

Machine learning can be mainly classified into two categories: Supervised learning and unsupervised learning. This paper select SVM as an example of supervised learning and K-means as an example of unsupervised learning to compare their performance. We get low accuracy values 47.02% for the former. Since unsupervised classification does not have a fix label, and using unsupervised learning may is not a good choice for our problem.

We then use K Nearest Neighbors (KNN) as a classifier to compare with SVM.

KNN aims at finding k nearest class:

$$p(d_{new}, c_i) = \sum_{k=1}^{m} Similarity(d_k, d_{new}) \times y(d_k, dc_i)(1)$$

$$(d_k, dc_i) = \begin{cases} 1 & d_k \ \ belongs \ \ to \ \ c_i \\ 0 & d_k \ \ not \ belongs \ to \ c_i \end{cases}$$

d_{new} stands for a new document need to be classified, d_k stands for the *kth* document of the corpus.

SVM aims at finding a hyper plane to classify samples:

$$svm(x) = \mathrm{sgn}\left\{\sum_{i=1}^{N} a_i y_i K(x_i \cdot x) + b\right\}(2)$$

Where $\mathrm{sgn}(\cdot)$ stands for sign function, and $k(\cdot)$ stands for the kernel function of SVM, a_i is determined by slack variables , y_i stands for the label of x_i , x stands for input text, b is determined by penalty factor.[17-20].

Using KNN [21], we achieve 63.15% precision, and the result suggests that KNN is also better than K-means, but is not as well as SVM which achieves 90.00% precision.

What's worse, KNN has lower efficiency than SVM when used to classify with big data. When the number of micro-blog needed to be classified grows from 1000 to 100000, the time consumed by KNN to finish this task is raised from 9.8 seconds to 524.67 seconds. However, the time consumed by SVM is always less than 1seconds.

Table 2. KNN classification confusion matrix

Manual / Classifier	Sick	Not Sick
Sick	17	40
Not Sick	2	55

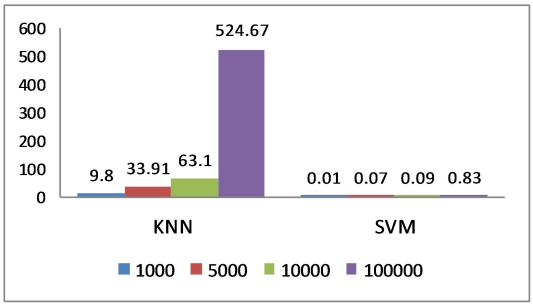

Fig 11. Time consumed by KNN and SVMLIGHT

D. Experiment

According to the experiment results above, we decide to use words as features, TFIDF as word weighting method, SVM as classifier, and we use SVM_LIGHT as a convenient tool for our experiments. This SVM is robust and includes algorithm for approximately training large transductive SVMs for big data set.

To prove this classifier is stable and reliable, this paper uses K-Fold cross validation for verification, and we use 5-fold for this test:

Table 3. 5-Fold classification experiment

Experiment / evaluation	1-kfold	2-kfold	3-kfold	4-kfold	5-fold
Precision	78.26%	90.00%	94.44%	86.21%	88.52%
Recall	94.71%	94.74%	89.47%	87.72%	94.74%
F1Measure	85.73%	92.31%	91.89%	86.96%	91.52%

From this experiment, we finally achieve 87.49% precision in average, 92.28% recall in average and 89.68% F1 measure in average. These prove that using our classification model can distinguish between 'sick micro-blog' and 'not sick micro-blog', and we use this model to our dataset from September 2013 to December 2013.

Table 4. Classification between September and December 2013

	Status No.	Sick status No.	Ratio
September	935646	63284	6.76366.9%
October	885436	60874	6.875031%
November	848685	57296	6.75115%
December	780890	55826	7.149022%

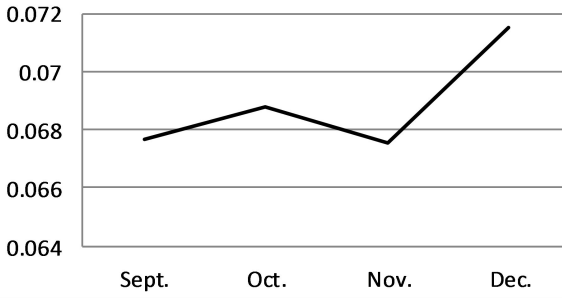

Fig 12. Classification result between Sept. and Dec. 2013

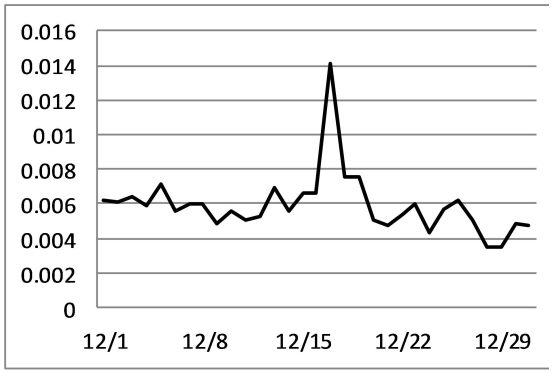

Fig 13. Classification result in December 2013

According to Figure 12, it can be seen that t the number of sick people has risen obviously, and this paper uses the same classification model to classify the micro-blogs day by day in December, and get the result as shown in Figure 13.

As shown in Figure 13, the ratio of 'sick micro-blog' rose from September 9th 2013, and peaked as 1.4% on September 17th 2013. Compared with the data from China Nation Influenza Center (CNIC), we predict outbreak time of flu 5 days earlier. CNIC's data suggests that seasonal influenza achieves a litter higher level from September 15th 2013, and rises to a much higher level in this month.

V. CONCLUSION

In this paper, we investigated how . to predict the trend of diseases in the real world based on Chinese social media. We combine social media data with spatio-temporal data, and successfully predict outbreak time of flu 5 days earlier than Chinese national influenza center. We also think that our method for processing Chinese social media data can be used in other related fields for Chinese big data analysis.

In the future, we need to take more considerations on spatio-temporal data, and investigate the flu's influence on individuals for a period time. Since most users of Sina micro-blogs are the generation born in 80s and 90s, we also need to obtain more data for other ages to get a more comprehensive result.

ACKNOWLEDGEMENT

This research is supported in part by the Fundamental Research Funds for the Central Universities of China No. 216-274213.

REFERENCES

[1] World Health Organization, 2003, Influenza fact sheet. http://www.who.int/mediacentre/factsheets/2003/fs211/en/

[2] IResearch,2010, In 2010 the global Internet users spend most of their time in social media. http://service.iresearch.cn/others//20101129/12857 3.shtml

[3] Infographic: The Growing Impact of Social Media. http://www.sociallyawareblog.com/2012/11/21/time-americans-spend-per-month-on-social-media-sites/

[4] Collier N, Son N T, Ngoc M N T, 2010, OMG U got flu? Analysis of shared health messages for bio-surveillance, Semantic Mining in Biomedicine.

[5] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2008, 457(7232): 1012-1014.

[6] Mangold W G, Faulds D J. Social media: The new hybrid element of the promotion mix. Business horizons, 2009, 52(4): 357-365.

[7] Kamel Boulos M N, Sanfilippo A P, Corley C D, et al. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine, 2010, 100(1): 16-23.

[8] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data. Nature, 2009, 457(7232): 1012-1014.

[9] Freifeld C C, Chunara R, Mekaru S R, et al. Participatory epidemiology: use of mobile phones for community-based health reporting. PLoS medicine, 2010, 7(12): e1000376.

[10] Sadilek A, Kautz H A, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data[C]//AAAI. 2012.

[11] Sadilek A, Kautz H A, Silenzio V. 2012, Modeling Spread of Disease from Social Interactions, ICWSM.

[12] Zhengyan Cui. Micro-blog classification based on semantics[J]. Modern Computer: the second half version,2010 (8): 18-20.

[13] Yang F, Liu Y, Yu X, et al, 2012, Automatic detection of rumor on Sina Weibo, Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM.

[14] Bao M, Yang N, Zhou L, et al. 2013, The Spatial Analysis of Weibo Check-in Data–The Case Study of Wuhan, Geo-Informatics in Resource Management and Sustainable Ecosystem. Springer Berlin Heidelberg, 480-491.

[15] Amati G, Van Rijsbergen C J. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 357-389.

[16] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management, 1988, 24(5): 513-523.

[17] Cortes C, Vapnik V. Support-vector networks. Machine learning, 1995, 20(3): 273-297.

[18] Joachims T. Svmlight: Support vector machine. SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund, 1999, 19(4).

[19] Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[20] Yang, N., Li, S., Liu, J., & Bian, F. 2014, Sensitivity of Support Vector Machine Classification to Various Training Features. TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(1), 286-291.

[21] Han E H S, Karypis G, Kumar V, 2001, Text categorization using weight adjusted k-nearest neighbor classification. Springer Berlin Heidelberg