An Intelligence System Based on Social Web Mining and Its Application in Health Care in Hong Kong

Kin Keung Lai[1,2]
[1] Management Sciences
College of Business, City University of Hong Kong, Hong Kong.
[2] International Business School
Shaanxi Normal University
Xian, China
mskklai@cityu.edu.hk

Juan Shi[1,2]
[1] Management Sciences
College of Business, City University of Hong Kong, Hong Kong.
[2] Management Science and Engineering
Xi'an Jiaotong University
Xian, China
juanshi4-c@my.cityu.edu.hks

Gang Chen
The 705th Research Institute
China Shipbuilding Industry Corporation
Xi'an, China
james.gang.chen@gmail.com

***Abstract*—In China, two systems—Chinese Information system for disease control and prevention (CDC) and the National Adverse Drug Reaction (ADR) Monitoring System--have been established for monitoring infectious diseases and for reporting ADR events. However, surveys have shown that nontrivial problems have affected the performance of these information systems adversely. Specifically, the absence of reasonable work and supervision systems, namely the systemic factors constitute the majority of the reasons for underperformance of these information systems. Also, not just some leaders but also quite a few doctors do not give sufficient importance to reports of infectious cases or ADR events, which is thought to be a waste of manpower and financial resources. Nowadays social media websites have attracted wide interest from both the academia and the industry, because of its very large user base, updating in real time, abundant incoming data and the fact that it offers prolific information about consumers' behavioral characteristics. In this proposal, we plan to address the aforementioned issues by implementing an Intelligence System based on Social Web Mining (ISSWM). ISSWM can help detect ADR events, assess the trends of infectious diseases, probe people's opinions towards some medical institutions or on healthcare related issues in real-time, at a cost that is extremely low compared with the existing systems.**

Keywords-big data; Web 2.0; social web mining; opinion mining; sentiment analysis; epidemics surveillance.

I. INTRODUCTION

The Chinese Information system for disease control and prevention (CDC) was established in China for control and prevention of infectious diseases in 2005. Medical and health institutions at the grass roots level are required to report cases of infectious diseases to this system, which helps the concerned department(s) to keep abreast of the situation and make timely decisions. Besides that, the National Adverse Drug Reaction (ADR) Monitoring System has also been launched to improve surveillance of ADR events. However, there are some nontrivial problems in these information reporting systems. For example, a survey, which investigates the reports of infectious cases via CDC in 17 medical institutions, has found some salient reasons that seem to be responsible for underperformance of CDC. First, reports using CDC have not been given sufficient importance by staff of these medical institutions. Second, there is a lack of effective work and monitoring mechanisms. Third, medical staff involved in reporting via CDC change frequently and lack adequate knowledge about infectious diseases [1]. As a result, missing reports rate is estimated at 11.76% and the rate of consistency between infectious diseases report cards and information input into the system is only 78.25%. Another survey about ADR has reached similar conclusions: both information reliability and timeliness leave much to be improved. From the above, it can be concluded that management factor and human factor are two main reasons of unsatisfactory performance of these information reporting systems. There is a long way to go and a lot of efforts are needed to address the problems of these systems.

Fortunately, we have entered the era of Web 2.0, which is characterized by high degree of public engagement and initiative. In China, the number of registered users of Sina Weibo, the biggest social platform, has reached 600 million and there are as many as 61.4 million daily active users [2]. It is worth noting that medical and health related data, generated on social media, constitute an important part of Big Data relevant to the medical area. Generated by users directly, they are constantly updated and reflect every aspect of people's lives [3]-[4]. More importantly, data on the social media are much easier to access via API compared with databases owned by the government. The rapid growth of online social networking sites and the public availability of data have made analysis of social media data easy and convenient [5]-[6].

This proposal envisages implementation of an intelligence system based on social web mining (ISSWM), which extracts insightful information from the unstructured or semi-structured data available on social media to answer questions related to people's main concerns in healthcare area at a certain moment. How is an epidemic spreading in

 ISBN: 978-1-61208-428-2

the country or within a certain area? What are people's specific expectations from some particular medical institutions? However, it needs to be pointed out that the purpose of ISSWM is not to substitute any existing authoritative information-gathering websites, such as CDC or the ADR monitoring system. Instead, ISSWM functions as an intelligence system complementary to the existing information systems. For example, it could serve as a plug-in for the existing systems. The advantages of ISSWM over existing information reporting systems, like CDC and ADR monitoring system, are as follows:

1. Timeliness and low cost. ISSWM can automatically collect, process and analyze data from social media websites in real time, using computer algorithms. These features of the proposed ISSWM have several implications. First, manual effort is greatly reduced and the process virtually costs nothing. Second, real-time data collection and analysis helps decision-making department(s) stay up to date with the latest situation. Third, ISSWM can effectively mitigate problems resulting from lack of reasonable regulation in CDC or the national ADR monitoring system and problems which arise from medical staff's lack of awareness regarding the importance of reporting infectious cases.

2. Tremendous user base generated by integrating users from multiple social networking websites. Since data can be collected from numerous users of multiple social media websites, ISSWM can efficiently extract information related to infectious diseases, ADR events and other related issues in healthcare sector. The very large number of users provides the much needed credibility to results of ISSWM.

3. Use of big data. During events of public interest, such as presidential debates, there are hundreds of thousands of tweets per minute [7]. Interactions on websites, such as Twitter and Sino Weibo, contain a huge amount of meaningful information and this presents some challenge to ISSWM, which is taken into account in design.

The rest of this paper is organized as follows. In Section 2, we discuss prior work and their limitations on exploiting social media data in healthcare area. Section 3 presents the framework of ISSWM. Section 4 concludes the paper.

II. LITERATURE REVIEW

Recently, a considerable amount of work has been dedicated to exploiting data on Twitter, most of which has focused on forecasting epidemics based on interactions on Twitter [8]-[12]. For example, the public sentiment with respect to H1N1 was tracked and actual disease activity was measured on the basis of information embedded in the Twitter stream [9]. Furthermore, Twitter data, such as friendship relationships [3], the proximity of users' geographical locations and users' mobility, are used to estimate the likelihood of whether a specific person is likely to get infected [10]-[11]. Influenza-related messages are identified by leveraging a document classifier and these messages achieve a correlation of 0.78 with the Centers for Disease Control and Prevention (CDC) statistics [12].

A recently emerging research area is to explore online reviews to understand patients' attitudes towards healthcare services. For example, it has been revealed that there is more than 80% agreement between patients' own quantitative ratings of care and those derived using sentiment analysis, from online free-text comments on different aspects of healthcare [13]. Reference [14] analyzed a corpus consisting of nearly 60,000 reviews with a probabilistic model of text. The output of the model was found to significantly correlate with state-level measures of healthcare.

After reviewing these literature, we find out that there are some limitations in the existing research. First, they focus on data collected from only one social media website, Twitter, as reported in [8]-[12], English National Health Service website in [13], and RageMDs in [14]. Due to a single source of data, it is doubtful whether the findings still hold in the context of the whole society; Twitter users are not necessarily representative of the wider population [15]. Second, existing research concentrates on solving one or two specific problems, such as detecting flu trend or tracking people's sentiment regarding H1N1. An integrated system is required to provide comprehensive knowledge about the current status and sentiment related to healthcare issues in the society. Third, most of prior studies have adopted a methodology called post-collection analysis [8, 10-12]. That is to say, the analysis depends on pre-retrieved tweets or a well-prepared corpus [13-14], and there is a time delay that cannot be negligible in their analysis results. Thus a real-time system is in required to ensure online analytical processing in a streaming fashion. In this way, the breakout of emergency events can be detected almost instaneously and the latest situations are reported in real time, given people's high engagement in social networking websites. Last but not the least, China has a much larger population compared with any other country, producing an unprecedented amount of social media data. However, healthcare related data on Chinese social networking websites are virtually untapped.

To bridge the abovementioned gaps, we propose an intelligence system based on social web mining (ISSWM). It collects data from multiple social networking websites, operates incessantly (24 hours*7 days) and delivers valuable results in real time. The structure of ISSWM is illustrated in Figure 1. Extensive explanation about ISSWM is presented in Section 3.

III. INTRODUCTION ABOUT ISSWM

The proposed ISSWM shall employ crawling techniques and a variety of machine learning algorithms, such as clustering, feature selection, k-nearest neighbors, regression and so on, to obtain insightful knowledge from social media websites. ISSWM mainly comprises four modules shown in Figure 1.

1. Data collecting and preprocessing module

As mentioned above, ISSWM incessantly collects data from several Chinese social networking websites, such as Sina Weibo [16], Tencent Weibo [17], Renren Net [18] and so on. Crawling technologies are also used as a complementary method to retrieve data, as there are a few limitations on obtaining tweets through API for some social networking websites [19]. Besides that, API access is not available for non-social networking websites [20]-[25]. It can

 ISBN: 978-1-61208-428-2

be expected that the raw tweets contain a lot of noisy data such as wrong spellings, punctuation, images, or tweets that have nothing to do with healthcare sector.

It is proposed to preprocess the raw data in the following way. First, regular expressions are used to delete numbers, punctuations, and non-Chinese characters. Second, stopwords, which are meaningless words, are filtered out. Third, we crawl healthcare websites, such as 120ask.com [20], mingpaohealth.com [22], etc., in order to get a corpus of healthcare related words. Inversed Term Frequency (IDF) [26] is used to evaluate how important a given word is. Words with higher IDF value are thought to be more related with healthcare sector. To further improve the accuracy of this corpus, a priori knowledge is used to filter out words which get higher IDF values but seldom appear in healthcare sector. After the healthcare corpus is established, raw tweets obtained from social networking websites are examined using this corpus. A tweet that does not share any word with this corpus is treated as non-relevant and is screened out.

Clean tweets are integrated, sorted and converted into a unified format, as illustrated in Table 1 and Table 2. Consistency needs to be checked as the tweets come from different data sources. Table 1 illustrates profile information about each user and Table 2 shows the format of information about each tweet. Relationship data such as each user's followers and friends are also kept in this step.

TABLE I. INFORMATION ABOUT USERS

Field	Description
id	ID of the user for whom to return results for table
screen_name	The screen name of the user for whom to return results for
statuses_count	The number of tweets issued by the user
description	Description about the user himself/herself
friends_count	The number of users this account is following
followers_count	The number of followers this account currently has
location	The user-defined location for this account's profile
time_zone	The time zone this user declares themselves within
…	…

TABLE II. INFORMATION ABOUT TWEET

Field	Description
id	The integer representation of the unique identifier for this Tweet
text	The actual text of the status update
created_at	Time when this Tweet was created
favorite_count	Indicates approximately how many times this Tweet has been "favorited" by Twitter users
retweet_count	Number of times this Tweet has been retweeted
place	Indicates that the tweet is associated a place
user	The user who posted this Tweet

2. Topic detection and classification module

We know that when the topic changes, keywords change correspondingly, since frequently used terms differ greatly across different areas. Latent dirichlet allocation (LDA) algorithm, which can represent each tweet as a random mixture of latent topics [27], is employed in this module to detect hot topics in social networking websites.

Suppose we would like to know what are the Top N topics attracting people's attention at a given point of time. First, we will retrieve tweets in the past month. After the preprocessing procedure (Module 1), we obtain clean tweets that focus on issues related to healthcare area, all of which are used to develop a LDA model, with the parameter "number of topics" set as N. Different weights are assigned on different topics that a given tweet can possibly be related to. The topic category on which a given tweet gets the highest weight is used to classify this tweet. In this way, all clean tweets are divided into different topic categories. Next, the N topics are sorted in terms of the number of tweets which belong to a particular topic. Finally, we get the Top N hot topics in healthcare area. The topic with a higher rank attracts more attention on social networking websites.

When a new unseen tweet arrives, it is transformed by using the established LDA model and is assigned to the topic category on which this tweet scores the highest. In this way, we identify which topic is acquiring more and more attention and, consequently, the trend of each of the Top N topics. Combing users' profile information and auxiliary information of a tweet, we can gain an even deeper understanding about the hot topics. For example, by exploiting the geographic information, we get to know the particular location where a specific topic induces a strong response among the citizens. With users' profile information, we may understand which group of people are involved in a particular topic. From time to time, we can check the content of the top N topics and get to know the change in the hot topics.

Like other hot topics, infectious diseases have their own key words, which can be obtained by a priori knowledge or learned by tweets on the internet. With these keywords, relevant tweets are extracted from multiple social networking websites and used to estimate the trend of some infectious diseases. In order to detect ADR events, a specialized corpus of medical-related words needs to be developed at the very beginning and any tweet which contains a relevant medical word or the name of a medicine in this corpus should be given special attention.

3. Module of opinion mining and sentiment analysis

After preprocessing (Module 1) and topic detection (Module 2), tweets are assigned to different topic classes. In order to grasp people's opinions or sentiments on some given topic, we use features like opinion words or phrases, negations, emoticons, syntactic dependency and so on to judge the sentiment underlying a tweet. Tweets which do not have these features are treated as neutral tweets and are discarded. Support vector machine (SVM) algorithm is employed to classify opinions into negative or positive and regression is used to identify the intensities of sentiments. As

 ISBN: 978-1-61208-428-2

both are supervised algorithm, labelled data are required to develop the model. We produce labelled data by labelling the polarities and intensities of the training tweets manually.

4. Opinion predicting module

Apart from classification of sentiments into positive or negative, we are interested in predicting people's opinions on new events. This problem has two dimensions. First, we want to forecast people's opinions on an upcoming event. Second, an event has occurred, but not everyone has publicly expressed his or her opinion toward this event. Therefore we would like to estimate the opinions of this silent group. We claim that people with similar characteristics tend to hold similar opinions toward the same event. Besides, people tend to hold similar opinions toward events of the same nature. The first hypothesis is validated in the following way.

First, we cluster people by characteristics, based on their profile information using clustering algorithms. Profile information generally includes age, occupation, gender and location and so on, as illustrated in Table 1, which can be extracted from tweets and people's own descriptions about themselves [28]. And then, we compare the similarity of people's attitudes towards a specified event, who are located in the same cluster with similarity of people's attitudes towards the same event, who are randomly chosen from a larger population. If the former similarity is significantly larger than the latter similarity in a statistical sense, the first hypothesis is validated. Following the same line, the second hypothesis could also be tested.

Based on the two validated hypotheses, we can use collaborative filtering which include memory-based collaborative filtering or model-based collaborative filtering to estimate people's opinions on new events.

IV. CONCLUSION

This paper envisages a new application of social web mining in medical and healthcare area. A system called ISSWM is proposed to exploit and analyze raw tweets from multiple social media websites. By ISSWM, we can predict the trends of infectious diseases, understand and estimate people's attitudes towards healthcare related issues, detect ADR events and stay informed about other healthcare related hot topics. ISSWM makes the best of social media websites and can deliver analysis results very efficiently at an extremely low cost.

ACKNOWLEDGMENT

The work described in the paper is financially supported by the Theme Based Research Grant, RGC of Hong Kong (TRS Project 8770001).

REFERENCES

[1] Y. He, Y. d. Liu, F. l. Ma, F. k. Xing, "Sample analysis in 17 medical institutions on the reports of infectious cases using Chinese Information system for disease control and prevention " Journal of Preventive Medicine of Chinese People's Liberation Army, no. 02, 2014, pp. 125-126.

[2] Sina (2014) Sina Weibo is the best crowd funding platform, take Dark Horse Games as an example. http://tech.sina.com.cn/i/2014-02-26/11479193540.shtml. Accessed 10 June 2015

[3] F. Griffiths, et,al, "Social networks–The future for health care delivery," Social science & medicine, vol. 75, no. 12, pp. 2233-2241, 2012.

[4] S. Brennan, A. Sadilek, H. Kautz, "Towards understanding global spread of disease from everyday interpersonal interactions," in Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013, pp. 2783-2789..

[5] C. McNab, "What social media offers to health professionals and citizens," Bulletin of the World Health Organization, vol. 87, no. 8, 2009, pp. 566-566.

[6] R. N. Rimal, and M. K. Lapinski, "Why health communication is important in public health," Bulletin of the World Health Organization, vol. 87, no. 4, 2009, pp. 247-247a.

[7] M. A. Russell, Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More: " O'Reilly Media, Inc.", 2013.

[8] R. Chunara, J. R. Andrews, J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak," The American Journal of Tropical Medicine and Hygiene, vol. 86, no. 1, 2012, pp. 39-45.

[9] A. Signorini, A. M. Segre, P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," PloS one, vol. 6, no. 5, 2011, pp. e19467.

[10] Sadilek A, Kautz HA, Silenzio V, Modeling Spread of Disease from Social Interactions. In: ICWSM, 2012.

[11] Sadilek A, Kautz HA, Silenzio V, Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In: AAAI, 2012.

[12] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in Proceedings of the first workshop on social media analytics, 2010, pp. 115-122..

[13] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, L. Donaldson, "Use of sentiment analysis for capturing patient experience from free-text comments posted online," Journal of medical Internet research, vol. 15, no. 11, 2013.

[14] B. C. Wallace, M. J. Paul, U. Sarkar, T. A. Trikalinos, M. Dredze, "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews," Journal of the American Medical Informatics Association, vol. 21, no. 6, 2014, pp. 1098-1103.

[15] D. King, D. Ramirez-Cano, F. Greaves, I. Vlaev, S. Beales, A. Darzi, "Twitter and the health reforms in the English National Health Service," Health policy, vol. 110, no. 2, 2013, pp. 291-297.

[16] Sina Weibo. "Sina Weibo API," 6th, July, 2015; http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5.

[17] Tencent. "Tencent Weibo API," 6th, July, 2015; http://dev.t.qq.com/.

[18] Renren. "Renren API," 6th, July, 2015; http://wiki.dev.renren.com/wiki/API.

[19] Sina Weibo. "Rate limiting on Sina Weibo API," 6th, July, 2015; http://open.weibo.com/wiki/Rate-limiting.

[20] 120ask. "120ask," 6th, July, 2015; http://www.120ask.com/.

 ISBN: 978-1-61208-428-2

[21] haodf. "haodf," 6th, July, 2015; http://www.haodf.com/.

[22] mingpaohealth. "mingpaohealth," 6th, July, 2015; http://www.mingpaohealth.com/cfm/Main.cfm.

[23] xywy. "xywy," 6th, July, 2015; http://www.xywy.com/.

[24] ewsos. "ewsos," 6th, July, 2015; http://www.ewsos.com/.

[25] 39net. "39net," 6th, July, 2015; http://www.39.net/.

[26] Wiki. "TF-IDF," 6th, July, 2015; https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[27] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, 2003, pp. 993-1022.

[28] J. Li, A. Ritter, E. Hovy, "Weakly supervised user profile extraction from twitter," ACL, Baltimore, 2014.

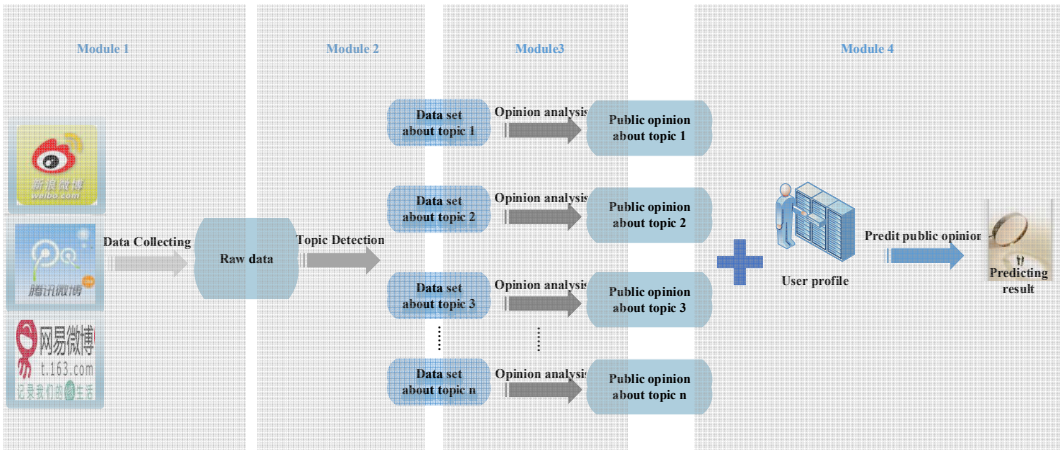

Figure 1. ISSWM for health care.