Exploring Health Topics in Chinese Social Media: An Analysis of Sina Weibo

Shiliang Wang, Michael J. Paul, Mark Dredze
Dept. of Computer Science and Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD 21218, USA
{wangshiliang,mpaul19,mdredze}@jhu.edu

Abstract

This paper seeks to identify and characterize health-related topics discussed on the Chinese microblogging website, Sina Weibo. We identified nearly 1 million messages containing health-related keywords, filtered from a dataset of 93 million messages spanning five years. We applied probabilistic topic models to this dataset and identified the prominent health topics. We show that a variety of health topics are discussed in Sina Weibo, and that four flu-related topics are correlated with monthly influenza case rates in China.

Introduction

Real-time user generated content on the web, epitomized by social media and in particular microblogs, are becoming an important data source to complement existing resources for disease surveillance (Brownstein, Freifeld, and Madoff 2009), behavioral medicine (Ayers, Althouse, and Dredze 2014), and public health (Dredze 2012). Most social media research in this area has focused on English-language data, with an emphasis on the United States (Culotta 2014) and United Kingdom (Lampos, De Bie, and Cristianini 2010), with some exceptions (Eiji Aramaki and Morita 2011; Chunara, Andrews, and Brownstein 2012). This paper presents, to the best of our knowledge, the first broad exploration of health-related content found in Chinese social media, through an analysis of nearly 100 million status updates from Sina Weibo, China's largest microblogging service.

As the world's most populous country, China has a central role in global public health. China was home to the 2003 SARS outbreak and the 2013 outbreak of H7N9 influenza, for which digital intelligence played an important role (Salathé et al. 2013). Rapid urbanization is posing a number of public health concerns, including increased accidents and injuries, healthcare disparities, and a growing disease burden due to changes in lifestyle and nutrition as well as increased air and water pollution (Gong et al. 2012). Tobacco use, a major global health concern, is prominent in China, accounting for a third of the world's smokers (Peto and Lopez 2001). All of these issues require up-to-date public health intelligence.

The aim of this research is to investigate which public health issues are discussed in Sina Weibo, hereafter referred to by its shorthand Weibo.[1] Based on prior work analyzing health topics in Twitter (Paul and Dredze 2011; Prier et al. 2011), we use probabilistic topic models to identify the prominent health topics in Weibo. We will show that Weibo contains a diverse set of health-related topics, including influenza, diet and nutrition, pregnancy and parenting, tobacco and alcohol, and air pollution. We show that influenza messages in Weibo are significantly correlated with influenza incidence rates in China (as high as r=0.73), suggesting that our data can be connected to real-world trends. We suggest future research directions given our findings.

Related Work

There is a large body of research on health in American social media, including surveillance of diseases such as influenza (Doan, Ohno-Machado, and Collier 2012; Broniatowski, Paul, and Dredze 2013), and the study of lifestyle factors such as physical activities (Yoon, Elhadad, and Bakken 2013) and tobacco use (Cobb et al. 2011). Far less research has been done on health in Chinese social media. Fung et al. (2013) analyzed Weibo messages pertaining to disease outbreaks including H7N9, with an emphasis on understanding public awareness and response. Hao et al. (2013) built models to predict the mental health status of Weibo users. There has also been research on influenza surveillance using Chinese web search data (Yuan et al. 2013), which is another digital data source that complements social media data. An overview of digital surveillance tools for China and elsewhere is given by Salathé et al. (2013). Our paper is the first comprehensive exploration of the variety of Chinese public health topics in social media.

Data

Weibo does not provide "streaming" API tools commonly used to obtain random samples over time from Twitter. We instead used the Weibo API[2] to query for all status updates from a given user, where users were crawled using a breadth-first search strategy beginning with random users with large

[1] "Weibo" (微博) is actually a general term – the Chinese word for "microblog" – but Sina Weibo is commonly called Weibo.

[2] http://http://open.weibo.com/wiki/API文档/en

(a)	感冒难受呀，喉咙发炎，不想说话！ The cold makes me so sick that my throat is sore, and I don't want to talk!
(b)	【治疗感冒的12种妙方】1热水泡脚；2生吃大葱；3盐水漱口；4冷水浴面；5按摩鼻沟: 12 treatments for curing cold: 1. Soak your feet in hot water; 2. Eat raw onions; 3. Gargle salt water; 4. Wash your face with cold water; 5. Massage the sides of your nose.

Figure 1: Two example Weibo messages. Most are from personal accounts (a), while some are from health organizations (b).

Year	All Data	Health Data
2009	40,837	805
2010	1,376,381	13,157
2011	7,758,806	67,250
2012	20,253,134	180,681
2013	63,789,097	658,280

Table 1: Number of messages per year in our collection.

numbers of followers. We obtained an average of 5 million statuses per day, and eventually collected 93 million status messages from approximately 1.6 million users. The messages span Nov. 2009 through Dec. 2013.

To filter for relevant health data, we followed the strategy of Paul and Dredze (2011, 2014) and first identified messages containing any of a set of health-related keywords. We scraped 598 disease names, 314 symptom terms, and 407 treatment terms (including medications) from a Chinese medical dictionary.[3] We additionally added the following keywords that were not part of the dictionary: 流感 (flu), 感冒 (cold), 医生 (doctor), 生病 (sick), 健康 (health), 节食 (diet), 锻炼 (exercise). After filtering by health keywords, we obtained a collection of approximately 920,000 health-related messages. Table 1 shows the yearly data volume.

Two examples of health messages from this collection are shown in Figure 1. Not all messages matching our keyword set are relevant to health. To understand the quality of our dataset, we randomly selected 100 messages for inspection. Two graduate student annotators labeled each message as relevant or not relevant to health, and the two annotators respectively found 57% and 59% of the messages to be relevant (κ=0.76). While we could further improve the relevance by filtering out false matches with supervised machine learning (Paul and Dredze 2014), the data quality was high enough for an initial exploration.

Our collection method poses new challenges compared to the sampling method of the Twitter streaming API. While the data covers 5 years, newer messages are favored (Table 1), where older messages will only appear for less active users, while active users have a smaller time period covered. Furthermore, collecting based on graph connectedness further biases the data, perhaps to specific regions or types of users. With this in mind, we emphasize that this is a purely exploratory study to investigate the types of topics that could be studied using Weibo. Furthermore, we will demonstrate promising results, even with this biased data, on a number of tasks. Our results motivate future work on obtaining unbiased samples from Weibo given the API restrictions.

[3] http://wubi.sogou.com/dict/cell.php?id=272

Topic Modeling Experiments

We employed probabilistic topic models to form a high-level understanding of the content of these 1 million health messages. We experimented with two topic models: Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) and the Ailment Topic Aspect Model (ATAM) (Paul and Dredze 2014). LDA models a document as a mixture of "topics" where each topic is characterized by a probability distribution over words. The estimated parameters of this model can be interpreted as clusters of words that tend to occur together in messages. ATAM is an extended model designed specifically for identifying health topics. ATAM combines standard LDA with a model for health topics (called "ailments") which are characterized by separate word distributions for symptoms, treatments, and general words. Both models were trained using Gibbs samplers.

Data Preparation

The Weibo messages were tokenized using MMSEG,[4] a word identification system for Mandarin Chinese text based on maximum matching algorithms (Chen and Liu 1992). We removed Chinese-language stop words[5] and words appearing infrequently in the data. Processed messages contained an average of 20 tokens, with a vocabulary size of 57,135.

Topic Discovery

Figure 2 shows topics discovered by LDA with 100 topics on all years of data. In addition to those shown in the table, we found topics describing healthcare and hospitalization, sleep issues (including 失眠 *(insomnia)*), muscle and joint pain (including 按摩 *(massage),* 肌肉 *(muscle),* 颈部 *(neck)*), common cold, skin disorders, exercise (including 运动 *(sports),* 散步 *(go for a walk)*), infant health, eye health (including 眼镜 *(glasses),* 近视 *(myopia)*), diet and weight loss, alcohol, and influenza.

Despite working well on Twitter, ATAM did not produce many coherent ailment clusters with Weibo, and in fact the model tended to allocate tokens to non-ailment topics (i.e. the LDA model embedded within ATAM). Perhaps the proportion of tokens matching our symptom and treatment dictionaries was too low for the sampler to assign these to ATAM's special symptom/treatment distributions. We will discuss LDA topics in the remainder of this paper.

External Validation

To evaluate how well the data corresponds to real-world trends, we considered the task of influenza detection, a com-

[4] http://technology.chtsai.org/mmseg/

[5] https://code.google.com/p/verymatch/

Tobacco		Nutrition		Pregnancy		Skin		Pollution	
丝瓜	(luffa)	食物	(food)	孕期	(pregnancy)	皮肤	(skin)	污染	(pollution)
戒烟	(quit smoking)	等	(wait)	孕妇	(pregnant woman)	肌肤	(skin)	环境	(environment)
抽烟	(smoking)	水果	(fruit)	胎儿	(fetus)	面膜	(facial mask)	日	(day)
肥皂	(soap)	维生	(vitamins)	妊娠	(gestation)	可以	(can)	中国	(China)
治疗	(treatment)	空腹	(empty stomach)	怀孕	(pregnant)	使用	(use)	月	(month)
慢性	(chronic)	营养	(nutrition)	妈妈	(mother)	毛孔	(pore)	空气	(air)
喉炎	(sore throat)	健康	(health)	周	(week)	保湿	(moisture)	年	(year)
可以	(can)	可以	(can)	发育	(develop)	芦荟	(aloe)	严重	(severe)
吸烟	(smoking)	消化	(digestion)	海带	(kelp)	护肤	(skin protection)	垃圾	(junk)
有害	(detrimental)	便秘	(constipation)	影响	(influence)	化妆	(makeup)	环保	(env. protection)

Figure 2: The ten most probable words for various health-related topics (bold headers assigned by annotators.)

ID	Words
2	感冒 (cold), 生病 (sick), 我在 (I'm in), 发烧 (fever), 难受 (uncomfortable), 头痛 (headache), 咳嗽 (cough), 抓狂 (crazy)
37	感冒 (cold), 注意 (notice), 天气 (weather), 保暖 (warm), 身体 (body), 冬季 (winter), 预防 (prevent), 大家 (everyone), 多吃 (eat more)
90	感冒 (cold), 生姜 (ginger), 红糖 (brown sugar), 开水 (boiling water), 睡觉 (sleeping), 继续 (keeping), 勺 (scoop), 蜂蜜 (honey)
95	疫苗 (vaccine), 流感 (flu), 儿子 (son), 病毒 (virus), 感染 (infect), 接种 (vaccinate), 医生 (doctor), 预防 (prevent)

Table 2: The top words for four flu-related topics. Topic 37 also includes "流感 (flu)" further down the top words list. Topic 90 also includes "出汗 (sweat)" and "寒气 (chill)".

mon application of digital disease detection. We identified four topics (shown in Table 2) that contain words describing flu or ILI symptoms. For each flu topic, we estimated the monthly influenza prevalence from Weibo by counting the number of messages containing the topic. The counts were normalized by the total number of messages from each month. We compared the Weibo-derived estimates to the monthly number of influenza cases obtained from the Chinese Center for Disease Control and Prevention (CCDC).[6] We measured the Pearson correlation between these two data sources during 2012 and 2013. (These two years account for over 90% of our data.) We excluded December 2013, since we began crawling near the start of this month.

Table 3 shows that across both years, topic 37 has the strongest correlation (r=0.56, p=0.005); this topic is shown alongside Chinese CDC data in Figure 3. Topic 2 has the highest correlation for 2012 (r=0.59) and 37 has the highest for 2013 (r=0.72). We obtained even higher correlations by counting the number of messages in the *union* of messages containing either one flu topic or another. In 2012, we obtain the highest correlation with messages containing topic 37 or 95 (r=0.65, p=0.02); for 2013, topics 37 and 90 (r=0.73, p=0.01). We find that all topics other than 37 are correlated with one year, but not the other (each topic has a moderate-to-good correlation for one year and no-to-low correlations for the other). This suggests a temporal covariate shift.

[6] http://www.chinacdc.cn/tjsj/

Year		Topic ID			
		2	37	90	95
2012	(n=12)	.59*	.50	-.0.05	.55
2013	(n=11)	.22	.72*	.46	.08
2012–13	(n=23)	.36	.56†	.16	.06

Table 3: Pearson correlation coefficients comparing CCDC influenza rates to different flu topics in different time intervals. Values are marked significant at $^{*}p<0.05$ and $^{\dagger}p<0.01$.

Discussion and Conclusion

In total, our topic model exploration discovered 16 distinct health issues, some of which were associated with multiple LDA topics (such as the four topics about influenza). One of the most interesting issues is pollution, because this was not a topic we discovered in our earlier work with Twitter (Paul and Dredze 2011; 2014), and is an important issue concerning both health and the environment. In addition to the pollution topic shown in Figure 2, LDA discovered a topic more specifically about air pollution, including the words "空气 (air)" and "呼吸 (breathe)".

The tobacco and alcohol topics were also not discovered in our earlier Twitter work. While these issues have both been studied in Twitter (Prier et al. 2011; Culotta 2012), they were not prominent enough to be discovered automatically during our untargeted topic modeling exploration. Because of their importance to public health and behaviorial medicine (Ayers, Althouse, and Dredze 2014), we believe these topics are worth pursuing in future research.

We noted other differences in the health topic composition of Weibo and Twitter.[7] There were several LDA topics describing foods, drinks, and herbs; these topics seem to be more common Weibo than Twitter. We also noticed more topics describing the health of infants and children. Even the common cold topic appears to describe sick children, as it includes "妈妈 (mother)" and "宝宝 (baby)" as top words. We did not find any topics describing mental health in Weibo, whereas with Twitter we have found topics describing depression and anxiety. However, neither "depression" nor "anxiety" appeared in our list of health-related keywords used to filter the data. These keywords should be included in future work, given the prevalence of

[7] Beyond health, see (Gao et al. 2012) for a comparitive analysis of the topics in Weibo and Twitter.

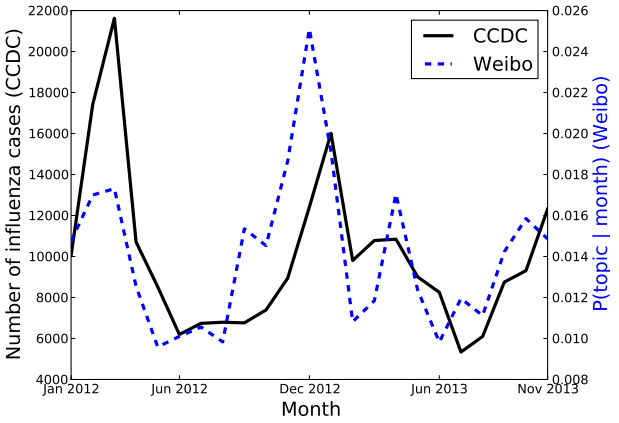

Figure 3: The prevalence of influenza across 2012–2013, as measured by the Chinese CDC compared to the proportion of messages containing the most-correlated flu topic.

mental disorders in China (Phillips et al. 2009) and recent work on mental health and Twitter (De Choudhury 2013; Coppersmith, Harman, and Dredze 2014).

Following up on any of these topics in depth requires a richer understanding of the data beyond the "bag of words" representation provided by topic models. It is not clear from the topic descriptions whether messages are from users sharing personal experiences, news stories, organizations providing advice or information, or advertisements. These distinctions matter when, for example, mining to learn about attitudes toward tobacco regulation or awareness of pollution effects. Our goal here is not to pursue these topics in depth, but to identify and characterize a broad variety of health issues that are discussed on Weibo. Our influenza experiments show the potential for Weibo as a public health source.

Acknowledgements

Qingjie Li assisted with relevance annotation. Jiefeng Zhai assisted with Chinese-to-English translation. Michael Paul is supported by a Microsoft Research PhD fellowship.

References

Ayers, J.; Althouse, B.; and Dredze, M. 2014. Could behavioral medicine lead the web data revolution? *JAMA*.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *JMLR*.

Broniatowski, D. A.; Paul, M. J.; and Dredze, M. 2013. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE* 8(12):e83672.

Brownstein, J.; Freifeld, C.; and Madoff, L. 2009. Digital disease detection – harnessing the web for public health surveillance. *N Engl J Med* 360(7256):2531–2157.

Chen, K., and Liu, S. 1992. Word identification for Mandarin Chinese sentences. In *COLING*.

Chunara, R.; Andrews, J.; and Brownstein, J. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg* 81.

Cobb, N.; Graham, A.; Byron, J.; Niaura, R.; and Abrams, D, B. 2011. Online social networks and smoking cessation: A scientific research agenda. *J Med Internet Res* 13.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Culotta, A. 2012. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*.

Culotta, A. 2014. Estimating county health statistics with Twitter. In *CHI*.

De Choudhury, M. 2013. Role of social media in tackling challenges in mental health. In *Workshop on Socially-Aware Multimedia*.

Doan, S.; Ohno-Machado, L.; and Collier, N. 2012. Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In *IEEE Conference on Healthcare Informatics, Imaging and Systems Biology*.

Dredze, M. 2012. How social media will change public health. *IEEE Intelligent Systems* 27(4):81–84.

Eiji Aramaki, S. M., and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP*.

Fung, I.; Fu, K.; Ying, Y.; Schaible, B.; Hao, Y.; Chan, C.; and Tse, Z. 2013. Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks. *Infect Dis Poverty* 2.

Gao, Q.; Abel, F.; Houben, G.-J.; and Yu, Y. 2012. A comparative study of users' microblogging behavior on sina weibo and twitter. In *20th International Conference on User Modeling, Adaptation, and Personalization*.

Gong, P.; Liang, S.; Carlton, E.; Jiang, Q.; Wu, J.; Wang, L.; and Remais, J. 2012. Urbanisation and health in China. *The Lancet* 379:843–852.

Hao, B.; Li, L.; Li, A.; and Zhu, T. 2013. Predicting mental health status on social media. In Rau, P., ed., *Cross-Cultural Design. Cultural Differences in Everyday Life*, volume 8024 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 101–110.

Lampos, V.; De Bie, T.; and Cristianini, N. 2010. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases* 599–602.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*.

Paul, M. J., and Dredze, M. 2014. Discovering health topics in social media using topic models.

Peto, R., and Lopez, A. 2001. Future worldwide health effects of current smoking patterns. In Koop, C., and Pearson, C.E. Schwarz, M. R., eds., *Critical Issues in Global Health*. Jossey-Bass.

Phillips, M.; Zhang, J.; Shi, Q.; Song, Z.; Ding, Z.; and et al. 2009. Prevalence, treatment, and associated disability of mental disorders in four provinces in China during 2001–05: an epidemiological survey. *The Lancet* 373:2041–2053.

Prier, K. W.; Smith, M. S.; Giraud-Carrier, C.; and Hanson, C. L. 2011. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *Conference on Social Computing, Behavioral-cultural Modeling and Prediction*.

Salathé, M.; Freifeld, C.; Mekaru, S.; Tomasulo, A.; and Brownstein, J. 2013. Influenza A (H7N9) and the importance of digital epidemiology. *N Engl J Med* 369:401–404.

Yoon, S.; Elhadad, N.; and Bakken, S. 2013. A practical approach for content mining of tweets. *Am J Prev Med* 45.

Yuan, Q.; Nsoesie, E. O.; Lv, B.; Peng, G.; Chunara, R.; and Brownstein, J. S. 2013. Monitoring influenza epidemics in China with search query from Baidu. *PLoS ONE* 8(5):e64323.