EXPLORING HEALTH TOPICS IN CHINESE SOCIAL MEDIA

SHILIANG WANG, MICHAEL PAUL, MARK DREDZE
JOHNS HOPKINS UNIVERSITY
GOALS OF THIS STUDY

• **Identify** a variety of different health issues that are prominently discussed in Chinese social media

  • Will use **topic models** to do this

• **Validate** utility and accuracy of health topics

  • Will compare **trends** to government surveillance data
    • Influenza
    • Air pollution (preliminary)
HEALTH IN SOCIAL MEDIA

- People publicly post a variety of information about their health through online social media
  - microblogs: Twitter, Sina Weibo

- People write about:
  - Acute illness (e.g. influenza)
  - Self medication (e.g. taking Nyquil)
  - Lifestyle/behaviors (e.g. going to the gym)
  - Alcohol, tobacco, drug use
  - Sleep habits
  - Mood

We can analyze messages on these topics to learn more
- “passive” approach to surveys
CHINESE SOCIAL MEDIA

• Sina Weibo
  • China’s most popular microblog
  • About 100 million active users
  • About 100 million messages per day

• Not extensively studied in this community
  • Especially relative to its popularity

• Many important public health issues in China
  • e.g. H7N9 influenza
RELATED WORK USING WEIBO

- Disease outbreaks
  - Fung; Fu; Ying; Schaible; Hao; Chan; Tse (2013)

- Mental health
  - Hao; Li; Li; Zhu (2013)

- Survey of digital epidemiology in China
  - Salatthe; Freifeld; Mekaru; Tomasulo; Brownstein (2013)

- Comparison to Twitter
  - Gao; Abel; Houben; Yu
• Weibo does not have “streams” like Twitter

• Breadth-first crawl:
  • Begin with a random user
  • Crawl all messages by that user
  • Repeat for each of the user’s followers

• We collected 93 million messages in Dec. 2013
  • messages span Nov 2009 – Dec 2013
DATA FILTERING

• Filtered for messages containing health-related keywords
  • 598 disease names
  • 314 symptom terms
  • 407 treatment terms

• Estimated that 58% are actually relevant to health
  • Two annotators labeled a sample of messages
  • Good enough for this exploratory study
DATA SET

- Nearly 1 million health-related messages:

<table>
<thead>
<tr>
<th>Year</th>
<th>All Data</th>
<th>Health Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>40,837</td>
<td>805</td>
</tr>
<tr>
<td>2010</td>
<td>1,376,381</td>
<td>13,157</td>
</tr>
<tr>
<td>2011</td>
<td>7,758,806</td>
<td>67,250</td>
</tr>
<tr>
<td>2012</td>
<td>20,253,134</td>
<td>180,681</td>
</tr>
<tr>
<td>2013</td>
<td>63,789,097</td>
<td>658,280</td>
</tr>
</tbody>
</table>
DATA EXPLORATION

- We used probabilistic **topic models** to identify prominent topics and themes in the health data

- **Unsupervised** clustering of words and messages into semantically coherent groups

- Used successfully in our earlier work with Twitter
  - ICWSM 2011; PLOS ONE 2014
TOPIC MODELING

- Latent Dirichlet Allocation (LDA) (Blei et al. 2003)
- Each document is a distribution over topics
- Each topic is a distribution over words
A jury found baseball star Roger Clemens not guilty on six charges against. Clemens was accused of lying to Congress in 2008 about his use of performance enhancing drugs.
TOPIC MODELING

football 0.03  
team 0.01 
hockey 0.01  
baseball 0.005 
...  

charge 0.02  
court 0.02 
police 0.015 
robbery 0.01 
...  

congress 0.02  
president 0.02 
election 0.015 
senate 0.01 
...  

Doc 1  

Doc 2  

Doc 3  

A jury found baseball star Roger Clemens not guilty on six charges against. Clemens was accused of lying to Congress in 2008 about his use of performance enhancing drugs.

Jury Finds Baseball Star Roger Clemens Not Guilty On All Counts
TOPICS DISCOVERED

• 16 distinct health issues:
  • Healthcare
  • Sleep issues
  • Muscle and joint pain
  • Common cold
  • Skin conditions
  • Skin health
  • Infant health
  • Eye health

  • Nutrition
  • Diet and weight loss
  • Exercise
  • Pregnancy
  • Pollution
  • Influenza
  • Alcohol use
  • Tobacco use
TOPICS DISCOVERED
COMPARISON TO TWITTER

• Some differences we noticed compared to our previous work with Twitter topic models:
  
  • Alcohol and tobacco use
    • Both have been studied in Twitter, but these weren’t discovered as topics by our methods in Twitter
  
  • Pollution
    • Two pollution topics in Weibo
  
  • Nutrition
    • Several topics about food, drink, and herbs
  
  • Infants and children
    • Multiple health topics
VALIDATION: INFLUENZA

- Compared the **temporal trend** of influenza-related topics to monthly data from the Chinese CDC
  - Four flu-like topics discovered by LDA

- Topics show moderate correlation with CCDC data:

<table>
<thead>
<tr>
<th>Year</th>
<th>Topic ID</th>
<th>2</th>
<th>37</th>
<th>90</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>(n=12)</td>
<td>.59*</td>
<td>.50</td>
<td>-.05</td>
<td>.55</td>
</tr>
<tr>
<td>2013</td>
<td>(n=11)</td>
<td>.22</td>
<td>.72*</td>
<td>.46</td>
<td>.08</td>
</tr>
<tr>
<td>2012–13</td>
<td>(n=23)</td>
<td>.36</td>
<td>.56†</td>
<td>.16</td>
<td>.06</td>
</tr>
</tbody>
</table>
VALIDATION: INFLUENZA
VALIDATION: AIR POLLUTION

• Compared the air pollution topic to government data on fine particle pollution (PM2.5) for 74 cities
  • Average daily value in 2013

• Correlation of .546

• Currently researching this topic more
LIMITATIONS

• Crawled data not a random sample
  • Presents difficulties for mining temporal trends

• Much of the data is noisy
  • But we’ve shown in past work that this can be cleaned up e.g. with supervised machine learning

• Concerns over censorship
  • Presumably a bigger problem for some topics more than others
CONCLUSION

• Many health topics are discussed in Weibo

• Early results show weibos are correlated with existing surveillance data

• Many health topics to potentially study in depth in future work
THANK YOU

• Acknowledgments:
  • Qingjie Li (annotation)
  • Jiefeng Zhai (translation)
  • Microsoft Research (PhD fellowship)