SOCIAL MEDIA AS A SENSOR OF AIR QUALITY AND PUBLIC RESPONSE IN CHINA

SHILIANG WANG, MICHAEL J. PAUL, MARK DREDZE
JOHNS HOPKINS UNIVERSITY

January 26, 2015 | AAAI Workshop on WWW for Public Health Intelligence

89°
Great pre-dawn, questionable everything else

the ONION
America's Finest News Source

China Vows To Begin Aggressively Falsifying Air Pollution Numbers

NEWS IN BRIEF · Environment · World · World Leaders · Barack Obama · News · ISSUE 50•45 · Nov 12, 2014

AIR POLLUTION IN CHINA

Air pollution is a public health issue
(also an environmental and political issue)

- Can directly cause many health effects
- Preventative perspective: what are people doing to protect themselves?

This is a rising and particularly severe problem in China

AIR POLLUTION IN CHINA

What are people saying about China's air quality on social media?

- Can social media detect pollution levels?
- Can we learn about health effects and behavioral response?

CHINESE SOCIAL MEDIA

Sina Weibo

- China's most popular microblog
- About 100 million active users
- About 100 million messages per day

新浪微博
weibo.com

- shorthand: "Weibo"

DATA COLLECTION

- Weibo's API does not provide "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

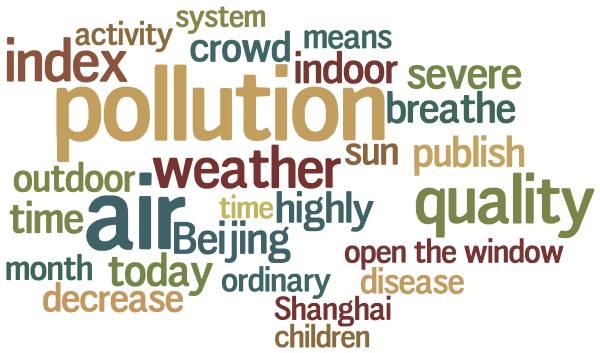
DATA COLLECTION

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

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

IDENTIFYING RELEVANT CONTENT

- Four keyword filters:
  - "pollution"
  - "air"
  - "breathe"
  - "cough"
- Two topic model filters (Latent Dirichlet Allocation):

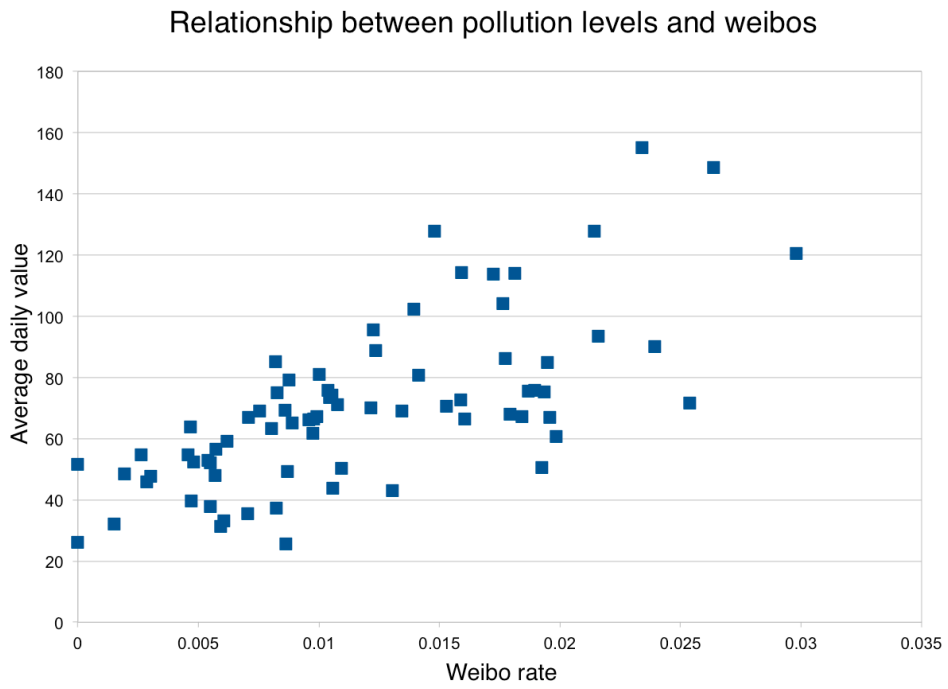


- Combined filters: topic + keyword

EXTERNAL VALIDATION

We compared the volume of messages with these filters to government-provided pollution levels

- The average daily value of fine particle (PM2.5) pollution levels across 74 cities in 2013

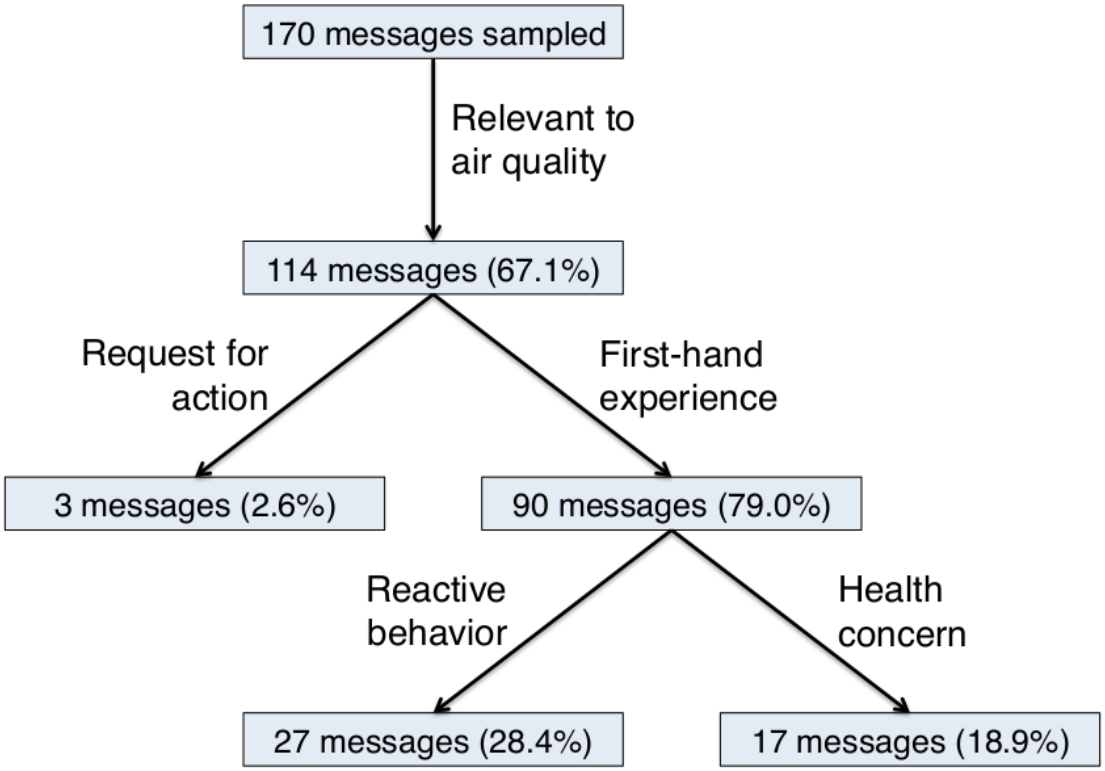


EXTERNAL VALIDATION

Filter	Correlation
AQ topic	.583
PO topic	.286
“pollution”	.610
“breathe”	.606
“cough”	.027
AQ+“air”	.623
AQ+“pollution”	**.703**

QUALITATIVE EXAMINATION

We annotated a small sample of topical messages with detailed codes



QUALITATIVE EXAMINATION

Not about pollution	累昏厥了。牢笼一般的机场巴士，传说中根本不叫花钱的物价，空气里的尿骚味以及灰蒙蒙的天。无论哪顿饭除了咖喱还是咖喱。 I was tired and fainting. The high price, the urine-scented air, and the heavy, gray day made the airport bus feel like a cage. Plus, every meal on the airport bus was curry.
About pollution, not a first-hand experience	老外说：这幅画表达的是污染程度的北京。PM爆表。 A foreigner said that this picture shows the serious pollution of Beijing. The PM value is too high.

QUALITATIVE EXAMINATION

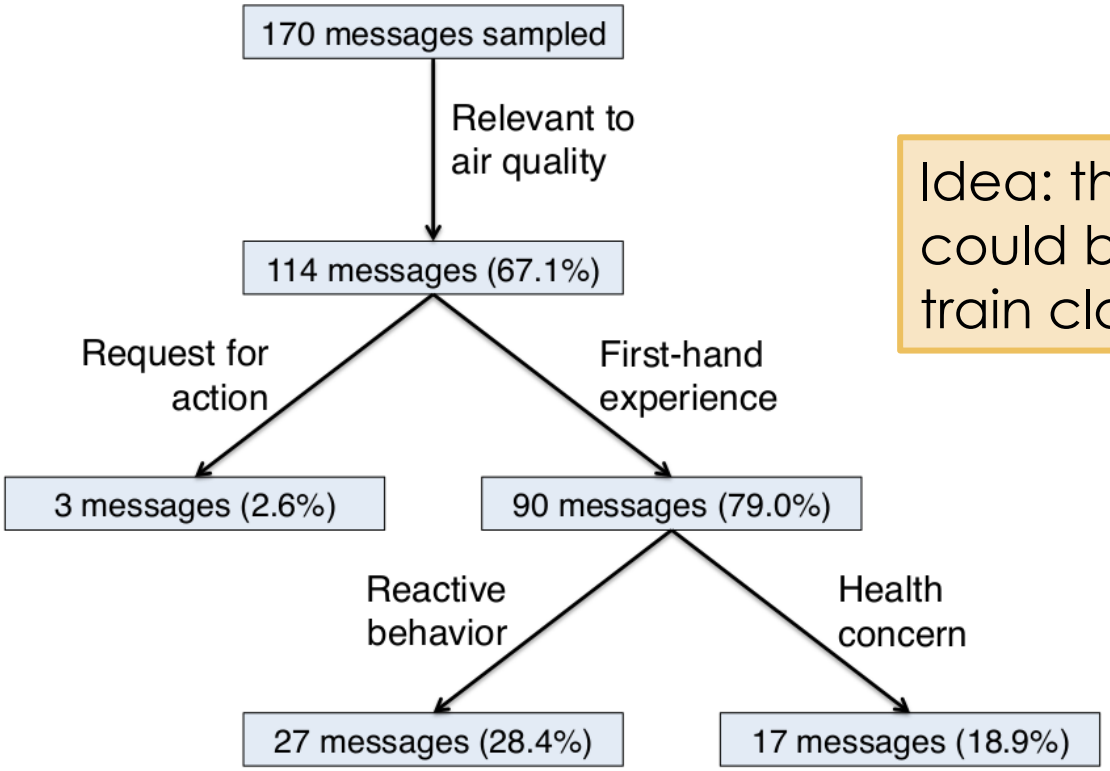
First-hand, reactive behavior	今晚想出去跑步,一查空气指数,还是轻度污染,在家避毒吧。 I want to go running this evening. However, it is lightly polluted based on the air pollution index, so I have to stay at home.
First-hand, health concern (+ reactive behavior)	三天前开始咳嗽。一定是北京污染的天气有关！以后出门戴口罩[生病]。 I start coughing three days ago. It must be caused by the pollution in Beijing! I will wear a mask when I go outside [sick].

QUALITATIVE EXAMINATION

Request for action	不能在空气质量重度污染时，才想起低碳行动！ Don't wait until the air has already been heavily polluted to start reducing carbon.

QUALITATIVE EXAMINATION

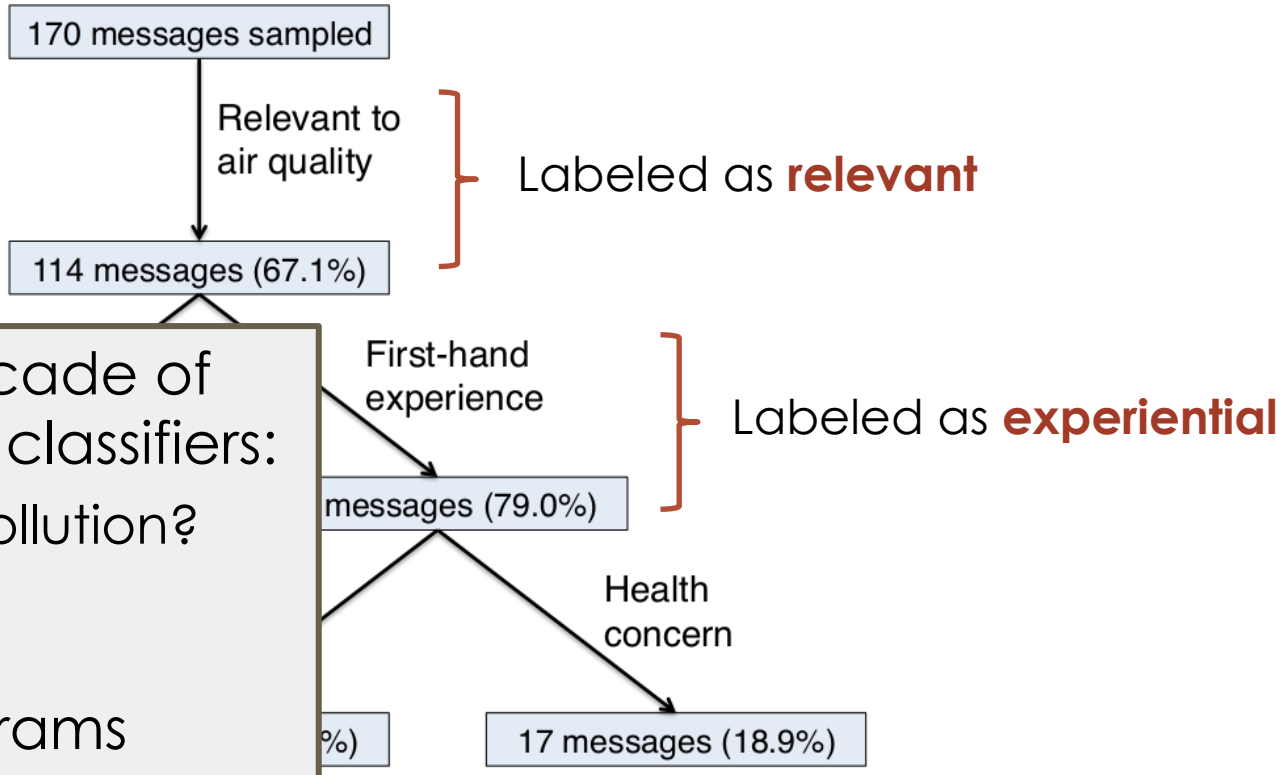
We annotated a small sample of topical messages with detailed codes



Idea: these codes could be used to train classifiers!

QUALITATIVE EXAMINATION

We annotated a small sample of topical messages with detailed codes



We trained a cascade of logistic regression classifiers:

1. **Relevant** to air pollution?
2. **Experiential**?

Features: 1, 2, 3-grams

EXTERNAL VALIDATION

Filter	Correlation
AQ topic	.583
PO topic	.286
“pollution”	.610
“breathe”	.606
“cough”	.027
AQ+“air”	.623
AQ+“pollution”	.703
Classifier	**.718**

RELATED WORK

- Mei S, Li H, Zhu X, Dyer CR. Inferring air pollution by sniffing social media. IEEE / ACM Int Conf Adv Soc Netw Anal Min. 2014.
  - Trains MRF to predict air pollution levels
  - Evaluates temporal trend
- Riga M, Karatzas K. Investigating the Relationship Between Social Media Content and Real-time Observations for Urban Air Quality and Public Health. Proc 4th Int Conf Web Intell Min Semant. 2014.
  - Looks instead Twitter (instead of Weibo)
  - Finds associations between text and health conditions

CONCLUSION

- Weibo is a rich source of interesting personal reports of health and behavior regarding air pollution
- Mentions of air quality are highly correlated with pollution volumes
  - Potential to act as auxiliary source of information, particularly for areas that do not have existing sensors

THANK YOU

With assistance from:

- Angie Chen
- Brian Schwartz