You Are What You Tweet: Analyzing Twitter for Public Health

Michael J. Paul and **Mark Dredze**
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{mpaul,mdredze}@cs.jhu.edu

Abstract

Analyzing user messages in social media can measure different population characteristics, including public health measures. For example, recent work has correlated Twitter messages with influenza rates in the United States; but this has largely been the extent of mining Twitter for public health. In this work, we consider a broader range of public health applications for Twitter. We apply the recently introduced Ailment Topic Aspect Model to over one and a half million health related tweets and discover mentions of over a dozen ailments, including allergies, obesity and insomnia. We introduce extensions to incorporate prior knowledge into this model and apply it to several tasks: tracking illnesses over times (syndromic surveillance), measuring behavioral risk factors, localizing illnesses by geographic region, and analyzing symptoms and medication usage. We show quantitative correlations with public health data and qualitative evaluations of model output. Our results suggest that Twitter has broad applicability for public health research.

Introduction

Twitter, Facebook and other social media encourage frequent user expressions of their thoughts, opinions and random details of their lives. Tweets and status updates range from important events to inane comments. Most messages contain little informational value but the aggregation of millions of messages can generate important knowledge. Several Twitter studies have demonstrated that aggregating millions of messages can provide valuable insights into a population. Barbosa and Feng (2010) classified tweets by sentiment, a first step towards measuring public opinion, such as political sentiment, which has been shown to track public political opinion and predict election results (Tumasjan et al. 2010; O'Connor et al. 2010). Eisenstein et al. (2010) studied lexical variations across geographic areas directly from tweets. Others have monitored the spread of news (Lerman and Ghosh 2010), detected the first mention of news events (Petrović, Osborne, and Lavrenko 2010), and monitored earthquakes (Sakaki, Okazaki, and Matsuo 2010).

Twitter users often publicly express personal information; messages like "I gots da flu" and "sick with this flu it's taking over my body ughhhh" are common. Knowing that a specific user has the flu may not be interesting, but millions of such messages can be revealing, such as tracking the influenza rate in the United Kingdom and United States (Lampos and Cristianini 2010; Culotta 2010b). While many studies have analyzed influenza rates and tracking in Twitter, these have largely been the limits of mining public health information from Twitter.

We believe Twitter can have a greater impact on public health informatics than just influenza tracking. A cursory examination of health related tweets reveals more detailed information: the message "Had to pop a Benadryl....allergies are the worst....ughh" indicates a user suffering from allergies and treating with Benadryl. "my g-ma is so cute when she's all doped up on Vicadin (sic.) for her foot" indicates Vicodin as a treatment for foot pain. Furthermore, tweets are not isolated events: they occur with specific times, locations, languages and users. Aggregating over millions of users could provide new tools for public health research.

This paper asks: what public health information can be learned from Twitter? While previous studies focused on specific questions (influenza rates) with specific models (keyword or statistical classifier for flu tweets), we ask an open question with a general model: the newly introduced Ailment Topic Aspect Model (ATAM) (Paul and Dredze 2011). Previous work introduced ATAM, created a data set and replicated influenza tracking results. This work improves the model using prior knowledge, and reports results for several new applications: geographic syndromic surveillance for multiple ailments (tracking illness over time and location), correlating behavioral risk factors with ailments, and analyzing correlations of symptoms and treatments with ailments. Our results include quantitative correlations with government data as well as qualitative evaluations of model output. To the best of our knowledge, this is the first work to use social media sources for a broad range of public health informatics on a range of ailments, rather than a narrow set of applications on one or two ailments.

Public Health Informatics and the Web

Syndromic surveillance, the monitoring of clinical syndromes that have significant impact on public health, impacts medical resource allocation, health policy and education. Many common diseases are continuously monitored by collecting data from health care facilities, a process known as sentinel surveillance. Resources limit surveillance, most especially for real time feedback. For this reason, the Web has become a source of syndromic surveillance, operating on a wider scale for a fraction of the cost. Google Flu Trends (Ginsberg et al. 2008) tracks the rate of influenza using query logs on a *daily* basis, up to 7 to 10 days faster than the Center for Disease Control and Prevention's (CDC) FluView (Carneiro and Mylonakis 2009). High correlations exist between Google queries and other diseases (Pelat et al. 2009), including "Lyme disease" (Seifter et al. 2010). These results fall under the area of infodemiology (Eysenbach 2009).

Similar results exist for Twitter, which can be a complimentary resource to query logs, but may also contain freely available and more detailed information; people write detailed messages for others to read. Lampos and Cristianini (2010) and Culotta (2010b) correlated tweets mentioning the flu and related symptoms with historical data. Similarly, Quincey and Kostkova (2010) collected tweets during the H1N1 pandemic for analysis. Ritterman, Osborne, and Klein (2009) combined prediction markets and Twitter to predict H1N1. More generally, Chew and Eysenbach (2010) evaluated Twitter as a means to monitor public perception of the 2009 H1N1 pandemic. Scanfeld, Scanfeld, and Larson (2010) evaluated the public understanding of antibiotics by manually reviewing Tweets that showed incorrect antibiotic use, e.g., using antibiotics for the flu.

The public health community is also considering how social media can be used to spread health information, with applications including risk communication and emergency response. Vance, Howe, and Dellavalle (2009) analyzed the pros and cons of using social media to spread public health information in young adults. Pros include low cost and rapid transmission, while cons included blind authorship, lack of source citation, and presentation of opinion as fact. Greene et al. (2010) studied how medical information is exchanged on Facebook, where groups specific to diseases share information, support, and engage patients in their diseases. Fernandez-Luque, Karlsen, and Bonander (2011) reviewed different approaches for extracting information from social web applications to personalize health care information. The model we use in this paper could be used to analyze tweets for health care personalization. Finally, the community is considering the larger impact of how social media can impact health care, where patients can "friend" doctors and constantly share information among thousands of friends (Hawn 2009; Jain 2009).

A Model for Diseases in Twitter

We apply the Ailment Topic Aspect Model (ATAM) (Paul and Dredze 2011) to created structured disease information from tweets that we use for public health metrics. To improve the quality of these metrics, we modify the model to incorporate outside information. We first summarize previous work on ATAM and then discuss our modifications.

Ailment Topic Aspect Model

Probabilistic topic models, such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), associate word tokens with latent *topics*. Documents are distributions over topics, and topics are distributions over words, oftentimes forming a semantically coherent word set. ATAM, which models how users express their illnesses and ailments in tweets, builds on the notion of topics. It assumes that for each health related tweet – we discuss how these are identified below – reflects a latent ailment a, such as flu, allergies, or cancer. Similar to a topic, an ailment a indexes a distribution over words ϕ_a. Additionally, an ailment maintains a distribution over symptoms and treatments, similar to the Topic Aspect Model (TAM) (Paul and Girju 2010). The latent variable $y \in \{0, 1, 2\}$ determines which aspect (general, symptom, treatment) generates each word. We follow Paul and Dredze (2011) and use a list of keyphrases to automatically identify possible symptoms and treatments (i.e. y is observed).[1] In addition to ailments, ATAM includes a more traditional topic model component – a topic z as a distribution over words and a document specific distribution θ over topics drawn from a Dirichlet distribution – since even in tweets about health topics, users also describe actions unrelated to this symptom-treatment structure, such as "sick today so playing video games." These topic distributions account for "playing video games." A switch variable $x \in \{0, 1\}$ (Binomialy distributed and parameterized by π) determines if a word is generated from an ailment dependent distribution $\phi_{a,y}$ or a non-ailment topic z. This idea is similar to the Twitter conversation+topic model (Ritter, Cherry, and Dolan 2010), where words in a tweet can depend either on a LDA-style topic model or a message-specific conversation act, though the model structure is different.

A Gibbs sampling algorithm learns ATAM's parameters. The collapsed Gibbs sampler marginalizes out the multinomials, requiring sampling for the variables a, z, x and ℓ from a distribution conditioned on the current values of all other variables. Details are beyond the scope of this summary. The topic Dirichlet hyper-parameter is optimized and other hyper-parameters are set manually. We used 8000 sampling iterations with $Z = 15$ and $A = 20$.

Model Extension: Incorporating Prior Knowledge

As with topic models, there is little control over what ailments ATAM discovers. While ATAM discovers meaningful ailments, there are numerous public health resources that could be used to improve model output. We use prior knowledge in the form of articles written about diseases, each an example of words associated with an ailment. We selected 20 disease articles likely to appear in Twitter data.[2]

[1] Lists are from the medical website `wrongdiagnosis.com`.

[2] From WebMD.com: Allergies, anxiety, asthma, back pain, breast cancer, COPD, depression, diabetes, ear infection, eye health, flu, foot injuries, heartburn, irritable bowel syndrome, migraine, obesity, oral health, skin health, sleep disorders.

We paired each article with an ATAM ailment and placed a prior distribution over the ailment's words based on the article. Following the LDA framework, the priors over our multinomial word distributions are defined by the Dirichlet distribution – the multinomials associated with each ailment are sampled from a different Dirichlet. The Dirichlet distribution can be defined with two parameters, a mean m, which can be thought of as the most likely multinomial to be drawn from this Dirichlet, and a precision s, which controls how much a sample multinomial vector can deviate from the mean – the lower the precision, the less influence the prior has over the inferred posterior. We set the mean m_a as the distribution over words in the articles for the ailment a. We optimize s to maximize data likelihood using the fixed-point iteration derived by Minka (2003) to update the precision given a fixed mean. The definition of the mean and the update rule for the precision are:

$$m_{a,w} = \frac{c_a^w + \lambda}{c_a^* + \lambda W} \tag{1}$$

$$s_a^{new} = s_a \frac{\sum_w m_{a,w}\Psi(n_a^w + s_a m_{a,w}) - m_{a,w}\Psi(s_a m_{a,w})}{\sum_w \Psi(n_a^* + s_a) - \Psi(s_a)} \tag{2}$$

The mean $m_{a,w}$ is a fixed constant, where c_a^w represents the count of word w in the article about ailment a. We use add-λ smoothing (with a small factor $\lambda = 0.01$) to ensure non-zero values for all words. The precision s_a is updated throughout the sampling process. We use a Gibbs EM procedure in which we perform 10 iterations of sampling with fixed Dirichlet parameters, then following Eq. (2) we update s based on the current state of the sampler, where n_a^w is the count of word w assigned to ailment distribution a. We ran this model with $Z = 20$ and $A = 20$ using the same hyperparameters as in the unsupervised setting. We call this new model ATAM+.

Evaluation

We begin with a description of data and a more traditional evaluation of model output. We start with a collection of over 2 billion tweets collected from May 2009 to October 2010 (O'Connor et al. 2010), from which we select health related tweets for ATAM training. Culotta (2010a) found simple keyword matching insufficient for filtering tweets (e.g., "I'm *sick* of this" and "justin beber ur so cool and i have beber *fever*"). Paul and Dredze (2011) use 5,128 labeled messages to train a high precision SVM binary classifier (0.90) to identify health related messages, which produced a corpus of 1.63 million English tweets. We removed punctuation, stop words, URLs and hashtags. ATAM (fully unsupervised) and ATAM+ (using WebMD articles as prior knowledge) were trained over these health messages.

Ailment output was annotated by two people not involved in running the models or aware of the WebMD articles. Each person labeled the 20 ailments with disease names of their own choosing or as incoherent based on the top 20 words for the general, symptom, and treatment distributions. For the 20 ailments, annotators agreed on labels for 17 ATAM ailments (of which 4 were incoherent). ATAM+ produced more clear and coherent ailments; annotators agreed on 15 ailments (all coherent) and no cluster was identified by both annotators as incoherent. Examples of seven of the fifteen ATAM+ ailments with annotator labels appear in Figure 1. The remaining ailments were respiratory illness, common cold, infections, physical injuries, headaches, exercise, skin problems and insomnia. To improve interpretability of this table, we include longer phrases in addition to single words. For each contiguous sequence of tokens in a tweet assigned to the same ailment distribution, we extracted a single phrase and included it in the ailment distributions. Several high frequency phrases appear in the top of the ailment distributions.

The discovered ailments qualitatively match their WebMD priors (first row of Figure 1), but the discovered ailments clusters often are more general. For example, the ailment with a "foot injuries" prior produced a general injuries ailment; "knee", "shoulder", and "arm" were top general words, "crutches" and "physical therapy" were top treatments. Likewise, the "breast cancer" prior resulted in general cancer; while "breast cancer" is the most frequent symptom, stomach, lung, and skin cancer all appear as top 10 symptom phrases. Additionally, we observe some confusions in the output between ailments with similar treatments or symptoms. For example, the "cancer" ailment also includes symptom phrases of "pneumonia" and "heart attack", and "heart surgery" is the top treatment phrase. These are all serious illnesses that appear within similar contexts (words like "pray" and "recovery") which caused the model to cluster them together. In general, this problem occurred more without the prior knowledge used in ATAM+. For example, allergies would appear with the "skin problems" ailment in ATAM, but ATAM+ kept allergic reactions (e.g. hives and rashes) separate from seasonal allergies.

Comparison to Public Health Articles

To evaluate the output interpretability, we compare the ailment distributions with distributions estimated from WebMD articles – the same 20 articles used as priors. To separately evaluate the symptom and treatment distributions, we filtered each article for symptoms and treatments using the keyphrase lists, then built ailment specific uni-gram language models. We then evaluate how well our ailment clusters match the WebMD articles by using these distributions to align the ATAM and ATAM+ ailments with the articles.

Each article was paired with its corresponding ailment in the model output, as labeled by the annotators – these pairings serve as our gold standard alignments – though not all articles had a corresponding ailment and vice versa. The Jenson-Shannon (JS) divergence was computed between the distributions for each ailment and the distributions for each WebMD disease. To account for ailments that contained very common symptoms and treatments, we measure the standard score: $\frac{x-\bar{x}}{\sigma}$ of the JS divergence, where $\bar{x}$ and σ are the mean and standard deviation of the divergence. This score was then used to define a ranking of articles for ailments, and ailments for articles. Ranking quality was measured as the percentage of articles/ailments such that the highest-ranked alignment was the correct alignment, as well as with mean reciprocal rank (MRR), i.e. the mean of $\frac{1}{\text{rank}}$, where rank is the position of the correct article/ailment in the list. Results

Ailment	Allergies	Depression	Aches/Pains	Cancer	Obesity	Flu	Dental
Prior	Allergies	Anxiety	Back Pain	Breast Cancer	Diabetes	Flu	Oral Health
Frequency	6.4%	5.8%	10.8%	8.0%	2.3%	8.1%	4.6%
General Words	allergies	help	body	cancer	blood	flu	meds
	stop	dont	head	pray	doctor	"swine flu"	killers
	eyes	body	need	mom	high	"flu shot"	dentist
	allergic	depression	hurts	shes	meds	dont	teeth
Symptoms	sneezing	pain	pain	pain	pressure	fever	pain
	cold	anxiety	aches	sad	"high blood	cold	toothache
	coughing	stomach	stomach	"breast cancer"	pressure"	"sore throat"	sore
Treatments	medicine	surgery	massage	surgery	hospital	hospital	braces
	benadryl	treatment	"hot bath"	hospital	diet	vaccine	surgery
	claritin	plastic	ibuprofen	"heart surgery"	exercise	medicine	antibiotics

Figure 1: Seven ailments discovered by ATAM+ represented by their most likely words, symptoms, treatments and the label independently assigned by two annotators. We also include the disease article used as a prior, and the discovered frequency of the ailment $p(a)$ in the 1.63 million tweets. Phrased extracted from the output are also shown.

Model	Total	Correct (S)	MRR (S)	Correct (T)	MRR (T)
Discovered Ailments to Articles					
ATAM	12	2	0.33	4	0.51
ATAM+	13	3	0.42	6	0.63
Articles to Discovered Ailments					
ATAM	12	0	0.23	0	0.36
ATAM+	13	1	0.32	5	0.54

Table 1: Results for aligning WebMD articles with discovered ailments using symptom (S) and treatment (T) models. Correct indicates the number of ailments with the correct answer at rank 1. Higher MRR means better matches between the articles and the discovered ailments.

(Table 1) show that ATAM+ outperforms ATAM, showing that the priors produced more coherent ailments. We also find that treatments are more consistent than symptoms, not an unexpected result given the commonality of symptoms for many of the ailments.

What Can be Learned from Twitter?

Given our evaluations suggesting that Twitter contains valuable information, we wonder: what public health information can be learned from Twitter? Using ATAM+, we explore several aspects of public health, including temporal and geographic impacts on medical wellbeing, and investigations into how the public treats illness.

Syndromic Surveillance

Public health data mining from Twitter has concentrated on syndromic surveillance, which tracks trends in medical conditions over time. Tracking influenza infections are especially conducive to syndromic surveillance since they are episodic and widespread. ATAM+ discovers many ailments, one of which is "flu." We measured the correlation between the probability of the flu ailment for each week (using the total tweets for that week) between mid-August 2009 to October 2010[3] with the influenza rate in the United States measured by the CDC, which collects and publishes detailed weekly statistics under FluView.[4] Results with ATAM+ (Figure 2) yielded a Pearson correlation coefficient of 0.958; ATAM obtained a correlation of 0.934. For comparison, we measured the correlation by normalizing using just the health related tweets, which considers the percentage of health related messages that discuss the flu. ATAM+ had a correlation of 0.966; ATAM had 0.881. For comparison, these numbers are in the range of the 0.95 of Culotta (2010a) and the 0.97 of Lampos and Cristianini (2010)[5]. However, an important difference is that these systems are trained to fit government data using regression, and were designed specifically to find flu tweets. In con-

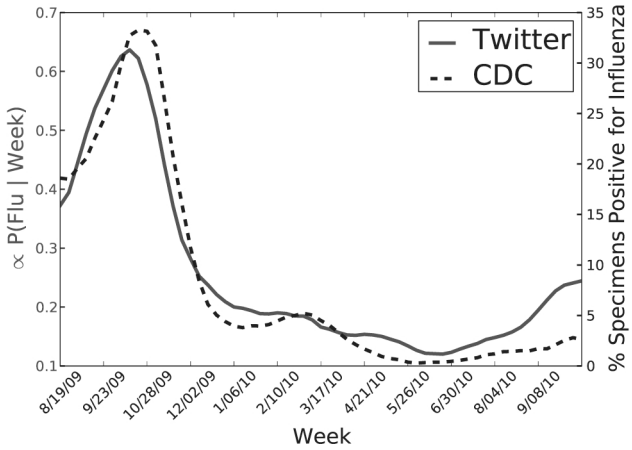

Figure 2: Influenza rate from August 2009 to October 2010 as measured by CDC FluView (measured as percentage of specimen's that tested positive for influenza) and ATAM+'s "flu" ailment (the normalized probability of the "flu" ailment given the week): correlation coefficient of 0.958.

[3]Earlier time periods contained gaps in the data.

[4]http://www.cdc.gov/flu/weekly/

[5]The former considers an unknown number of tweets from September 2009 to May 2010; the latter considers approximately 26 million UK-based messages from the second half of 2009.

trast, our approach discovers many ailments without labeled tweets, of which one is "flu".

Finally, since Google Flu Trends (Ginsberg et al. 2008) tracks the rate of influenza using query logs, we compare our Twitter rates directly with their reported rates. We obtained the Google Flu Trends data[6] for each week in the same time period and measured the correlation with our Twitter rates. When normalizing by all tweets, ATAM+ obtains 0.968 and ATAM 0.935. Normalizing by only health related tweets yields 0.974 for ATAM+ and 0.872 for ATAM.

Geographic Behavioral Risk Factors

Most work on automated syndromic surveillance focuses on national health trends. However, syndromic surveillance, which often focuses on seasonal diseases, is only one of a myriad of measures used to track population wellbeing, like tracking behavioral risk factors, such as the rate of cancer or obesity. Sentinel surveillance collects health statistics for chronic diseases or risk factors, often based on geography.

We investigated extracting geographically linked health statistics from Twitter. For example: are users in some geographic areas more likely to exercise than others? We formulated a list of questions based on the behavioral risk factor surveillance system, run by the National Center for Chronic Disease Prevention and Health Promotion at the CDC. For each statistic that could potentially be measured with one or more ATAM ailments, we measured the Pearson correlation coefficient between the ailments discovered in each US state with the state's risk factor rate. For example, we measured reported responses to the question "During the past month, did you participate in any physical activities?" with the exercise ailment for each state. This data was collected through phone interviews of over 350,000 US adults in 2009.[7]

We assigned a US state to each tweet according to user location by extracting this information from the location field, part of a user's profile. We searched each location field for the name of a state (California, Wisconsin, etc.), or a state abbreviation (CA, WI). The abbreviation had to be either uppercase, or appear in the format "——, ca". These strict rules reduced the number of false positives.[8] We extracted location information from 12% of the health related tweets (196,000 tweets). We computed the number of each ailment occurrences per capita: occurrences divided by the total number of tweets per state.

Table 2 shows the correlations between each risk factor and the related ailments measured for each state that had at least 500 tweets (42 states in total.) The strongest correlation was for tobacco use (proportion of residents in each state who are smokers) with the cancer ailment. The cancer ailment also had a high correlation with the percentage of people who have had heart attacks since some heart-related ailments were grouped into the cancer cluster. Other interesting correlations include a significant negative correlation between exercise and the frequency of posting any ailments, suggesting that Twitter users are less likely to become sick in states where people exercise. Similarly, there is a negative correlation between the rate of health care coverage in a state and the likelihood of posting tweets about diseases.

Risk Factor	Ailments	Correlation ATAM+	Correlation ATAM
Asthma	Allergies	0.241	0.195
Diabetes	Obesity	0.073	0.203
Exercise	All ailments	-0.352	-0.394
Exercise	Exercise	0.140	–
Exercise	Obesity	-0.201	-0.248
Health Care Coverage	All ailments	-0.253	-0.319
Heart attack	Obesity	0.244	0.341
Heart attack	Cancer	0.613	0.291
Obesity	Obesity	0.280	0.203
Obesity	Exercise	-0.267	–
Tobacco use	Cancer	0.648	0.320

Table 2: Correlations between the rate of a behavioral risk factor in a US state using CDC data with the measured rate of an ATAM ailment in that state. For example, states with higher obesity rates had more obesity tweets, and states with higher smoking rates had more cancer tweets. Note: ATAM+ discovered an exercise ailment; ATAM did not.

Geographic Syndromic Surveillance

So far we have demonstrated the ability to mine public health information both over time (syndromic surveillance for influenza) and by geographic region (behavioral risk factors.) We now seek to combine these to track an ailment over time and geography. For this experiment, we select seasonal allergies, an ailment that is correlated both with time and location; allergy season ranges over different dates by region. We computed the (normalized) rate of the allergy ailment by state and by month and plotted four months: February, April, June, and August (Figure 3). Shading was computed from the mean and standard deviation of all the states' allergy rates (over the three months) so that the color scale (ranging from lighter to darker) indicates rates that were more than a standard deviation below, half a standard deviation below, within half a standard deviation, more than half a standard deviation above, and more than a full standard deviation above the mean. States with less than 50 tweets in a month were excluded (shaded diagonally.) We also show the top 10 words w most predictive of the allergy ailment a and a given month m, ranked according to $p(a, m|w)$.

We observe several known patterns of allergies.[9] The primary allergy season begins in February or March and runs through May or June, beginning earliest in the South and the West. Observe that these areas have higher rates than the rest of the nation in February, while high rates have appeared in the rest of the country by April, the peak of the season. The

[6]Data Source: Google Flu Trends (www.google.org/flutrends)

[7]http://apps.nccd.cdc.gov/gisbrfss/

[8]Many self-reported locations are nonsensical (e.x.: "Fighting Dragons" or "under ur bed") (Hong 2011), but strict patterns seemed to produce good matches. Twitter provides geocodes, but this relatively new feature was missing from much of our data. In the future as geocodes increase, we can use them instead.

[9]Seasonal allergy information is taken from the WedMB article on allergies and geography: http://www.webmd.com/allergies/allergy-relief-10/worst-allergy-cities.

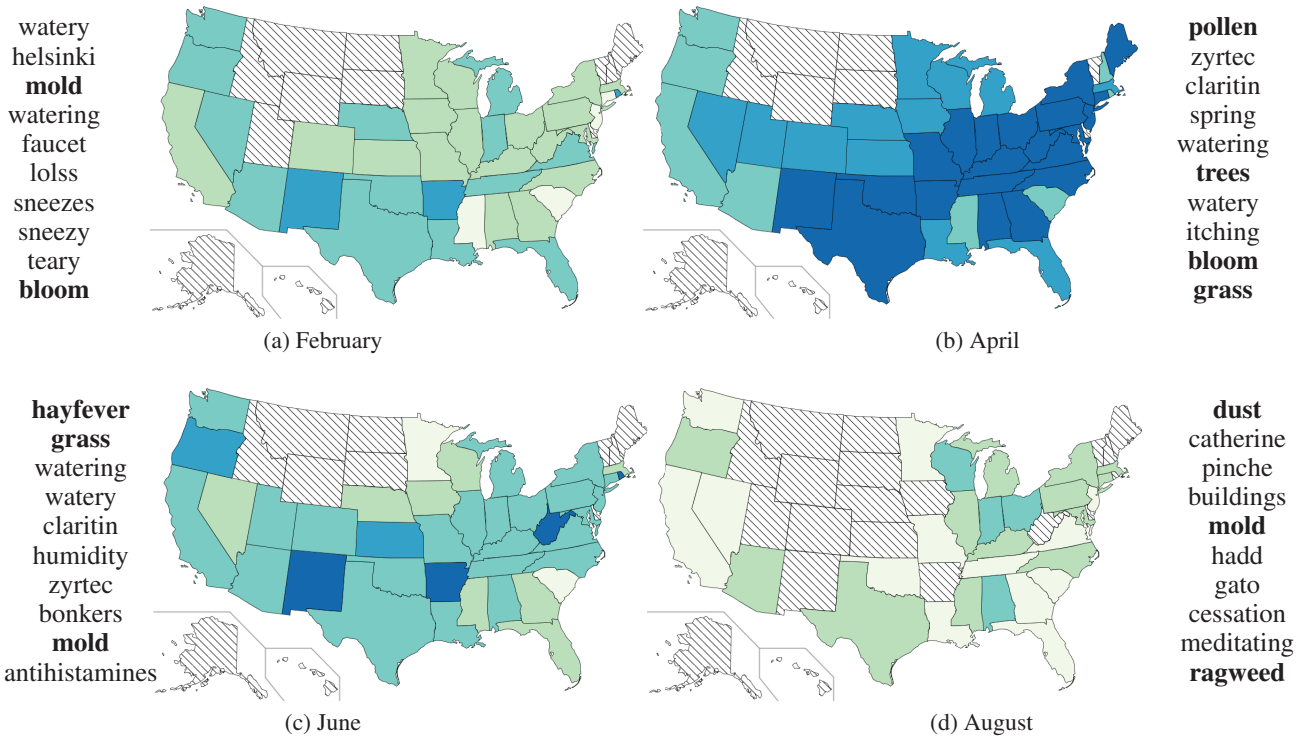

Figure 3: Rates of allergies discovered by ATAM+ by state for four months in 2010. Darker shading indicates higher allergy rates; diagonal shaded states lacked sufficient data. Notice that allergy rates peak in the middle of the allergy season (April), and that in colder weather (February) they are more prominent in warmer climates (West and South). Next to each month are the top 10 words w most predictive of the allergy ailment a during the month m ($p(a, m|w)$). Bolded words are those especially recognized as being associated with particular allergy seasons, e.g. ragweed begins to pollinate in mid-August, tree pollen season is in full effect in April, and June is often the worst month for hay fever (`http://www.gsk.com/infocus/hay-fever.htm`).

primary allergen in the spring is tree pollen, indicated by the presence of "pollen" and "trees" in April. Grass allergies peak in May and June; observe that "grass" is one of the top words in June. Weed season begins in August, the main contributor of which is ragweed, which is known to be especially prevalent in the Midwest. Notice that "ragweed" is associated with August, and that many Midwestern states have the highest rate during this month.

We seek quantitative validation of trends by time and location. Returning to the task of influenza tracking, the CDC provides infection rates by state.[10] We compute the influenza rates using the flu ailment as before, but now compute separate rates for each state. We measured the rates for the 2009-2010 influenza season (12 months from September). The CDC data uses a numeric value (integer between 1 and 10) indicating the level of flu activity for each week in each state. We ignored state/week pairs where the total number of tweets was less than 50, reducing the number of data points to 39% of the total, yielding 1057 data points. ATAM+ obtained a correlation of 0.66, and ATAM yielded 0.23. These correlations demonstrate the feasibility of syndromic surveillance within specific geographic regions.

Word	Ent.	Most Common Ailments
vomiting	2.19	Flu (23%), Aches (16%), Insomnia (12%)
burn	2.02	Skin (36%), Aches (17%), Headache (2%)
chill	1.95	Headache (28%), Insomnia (18%), Flu (12%)
fever	1.46	Flu (50%), Cold (24%), Infection (11%)
pimples	0.72	Skin (84%), Depression (5%)
fractured	0.69	Physical injuries (82%), Cancer (12%)
toothache	0.61	Dental (83%), Insomnia (9%), Aches (6%)
headache	0.56	Headache (75%), Insomnia (25%)
tumor	0.22	Cancer (96%)
mood	0.20	Depression (96%), Obesity (4%)

Table 3: Common symptom words, their most commonly associated ailments, and the entropy of their distribution over ailments. Entropy values ranged from 0 to 2.47.

Analyzing Symptoms and Medications

Discovered ailments include both milder, acute illness (influenza, allergies, etc.) as well as more common chronic conditions (obesity and insomnia). For many of these ailments, sufferers often do not visit their doctor, instead managing the illness on their own. Flu patients often stay home, take medication and eat soup. Obese people put themselves on a diet and exercise regimen. Insomnia suffers resort to changing their sleeping environment or sleeping aids. Since people are not seeing health care providers, their illness, symptoms

[10] `http://gis.cdc.gov/grasp/fluview/main.html`

Word	#	Ent.	Most Common Ailments
Pain Relief Medication			
tylenol	1807	1.57	HA (39%), IN (30%), Cold (9%)
ibuprofen	1125	1.54	HA (37%), DN (21%), Aches (17%)
advil	1093	1.08	HA (61%), Cold (6%), DN (5%)
aspirin	885	1.04	HA (69%), IN (10%), Aches (10%)
vicodin	505	1.33	DN (61%), Injuries (11%), HA (10%)
codeine	406	1.94	Cold (25%), DN (19%), HA (17%)
morphine	206	1.17	DN (59%), Infection (22%), Aches (9%)
aleve	183	1.10	HA (62%), IN (15%), DN (14%)
Allergy Medication			
benadryl	871	1.24	Allergies (64%), Skin (13%), IN (12%)
claritin	417	0.54	Allergies (88%), HA (5%)
zyrtec	386	0.49	Allergies (90%)
sudafed	298	1.61	Allergies (39%), Cold (21%), HA (20%)

Table 4: Treatment words (medications) for pain relief and allergies sorted by frequency (#), their most commonly associated ailments, and the entropy of their distribution over ailments. Entropy values ranged from 0.02 to 2.26. Common ailments discovered for pain relief medications are headaches (HA), insomnia (IN) and dental (DN).

and chosen treatments go unreported. Obtaining these statistics requires extensive polling of a large population. Twitter provides an opportunity to quickly and easily measure these statistics. Scanfeld, Scanfeld, and Larson (2010) consider one such application, in which they studied how people used antibiotics by examining tweets.

For each symptom and treatment w that appeared more than 50 times in the health related messages, a total of 413 symptoms and 113 treatments, we computed the distribution over ailments using ATAM+: $P(a|w)$. Low entropy treatments indicate specialized use, such as "sunscreen" (treatment) and "pancreatic" (symptom). High entropy indicates common use, such as "medicine" (treatment) and "painful" (symptom). We provide two sets of results based on this data. Common symptoms, their entropy, and most common associated ailments appear in Table 3. These distributions correspond to common knowledge about symptoms: fevers are associated with the flu, the common cold and infections, and toothaches are linked to dental problems, insomnia, and aches. Other terms showed interesting correlations, such as "pimples" and depression, and "mood" (as in bad mood or mood swings) with obesity.

To evaluate treatments, we selected two common groups of medications: pain relief and allergy medication. We compare several common medications within these groups in Table 4. The most common ailment for the pain relievers is headaches. The three strongest medications, codeine, morphine and vicodin, are all used for more serious ailments, including dental problems and infections. Tylenol (acetaminophen) is perhaps the most popular pain reliever on the market, and is the most commonly mentioned pain reliever. Additionally, ibuprofen is used to treat aches as it is an anti-inflammatory. The allergy medications are all primarily used to treat allergies. Benadryl is a more generic cold/allergy medication that causes drowsiness, thus it is used for insomnia in addition to allergies. Similarly, Sudafed is marketed for sinus headaches, so it appears with the common cold and headaches. As other examples of treatments, we found "penicillin" (entropy 1.64) most commonly appearing with infections (45%) and "acupuncture" (entropy 1.45) most commonly appearing with depression (36%) and physical injuries (33%). Finally, similar to Scanfeld, Scanfeld, and Larson (2010), we found that while the most common usage of antibiotics was correctly for infection (68%) and dental (9%) (patients whose wisdom teeth are removed often receive antibiotics to prevent infection), there were incorrect uses as well: common cold (9%) and flu (3%). These findings largely confirm the known properties and intended usages of common medications.

The Limits of Twitter

We have investigated a variety of public health data that can be automatically extracted from Twitter. These results suggest that public health officials and researchers can both replace current more expensive and time consuming methods and expand the range of what can be easily measured, including both new sources and types of information. However, Twitter cannot provide all answers, and it may not be reliable for certain types of information. We describe some of the limitations we encountered in this study.

We focused on population level metrics, but we had wanted to consider analyses on the scale of individual users. For example, how often does an allergy sufferer have serious allergy problems? What is the recurrence rate of the flu? Are certain people more likely to have repeated physical injuries? Are various ailments correlated with different lifestyles? Additionally, we sought to compute statistics about the course of acute illnesses. How long do influenza symptoms last? Which medications have the largest impact on symptoms? All of these statistics require that a single user post multiple tweets, and possibly multiple updates over the course of an illness. However, in our health related messages, 71% of users have only a single tweet and 97% have 5 or less, insufficient for computing user level statistics.

Our geographic analysis was fairly coarse, providing statistics on the state level. More sophisticated geocoding techniques could provide finer grained location information, allowing for the detailed analysis as in Eisenstein et al. (2010). However, we had too few tweets to consider finer grained statistics, e.g. on the county level. We failed to obtain even sufficient messages for each state. Additionally, we observed that some tweets, particularly those about more serious illnesses like cancer, were in regards to family members rather than the user. This could throw off geographic statistics if the subject of a tweet lives in a different location than the author of the tweet. This issue could perhaps be solved by filtering out tweets that are not about the user, for example by using a classifier that is trained for this task.

The expression "more data is better data" holds here as well. Our best results were obtained for influenza tracking (high correlation with CDC data) which represented 8% of the 1.63 million tweets. The above limitations are all inherently data limitations, so more data suggests that some of these goals may be achievable. The initial data set contained 2 billion tweets, which ranged from 1 millions message a day (mid 2009) to close to 6 million messages a day (late

2010.) While 2 billion is a seemingly massive number, it represents a fraction of the total messages available. Twitter reports that users now generate 50 million messages per day.[11] Perhaps access to this quantity of data will enable many of these statistics. Furthermore, the usage of mobile devices and geotagging messages has greatly increased, suggesting that more detailed geographical analyses are possible.

Finally, Twitter user demographics will limit research. Twitter users tend to be younger (nearly half are under 35) and only 2% are 65 or older.[12] Twitter is still US centric: as of June 2009, 62% of Twitter users were in the US.[13] These impose limitations on mining health information about senior citizens, or other countries. However, increasing Twitter usage may yield sufficient data. Future work will clarify what are true limits of this research direction.

Future Directions

Our exploratory study of using Twitter to mine public health information focused on producing data that correlates with public health metrics and knowledge. Twitter clearly contains many different types of information of value to public health research on many different ailments. The next step is to consider what new information can be learned by studying Twitter, potentially supporting new health informatics hypotheses. We plan to consider more specific applications with the goal of learning new things from Twitter.

Acknowledgements

We thank Adam Teichert for annotation assistance, and the anonymous reviewers for their feedback. The first author is supported by a NSF Graduate Research Fellowship.

References

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING*.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* 3.

Carneiro, H., and Mylonakis, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis* 49(10):1557–64.

Chew, C., and Eysenbach, G. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE* 5(11):e14118.

Culotta, A. 2010a. Detecting influenza epidemics by analyzing twitter messages. arXiv:1007.4748v1 [cs.IR].

Culotta, A. 2010b. Towards detecting influenza epidemics by analyzing twitter messages. In *KDD Workshop on Social Media Analytics*.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Empirical Natural Language Processing Conference (EMNLP)*.

Eysenbach, G. 2009. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J Med Internet Res* 11(1):e11.

[11] http://blog.twitter.com/2010/02/measuring-tweets.html

[12] pewinternet.org/Reports/2009/Twitter-and-status-updating.aspx

[13] http://www.sysomos.com/insidetwitter/

Fernandez-Luque, L.; Karlsen, R.; and Bonander, J. 2011. Review of extracting information from the social web for health personalization. *Journal of Medical Internet Research* 13(1).

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Greene, J.; Choudhry, N.; Kilabuk, E.; and Shrank, W. 2010. Online social networking by patients with diabetes: A qualitative evaluation of communication with facebook. *Journal of General Internal Medicine* 1–6. 10.1007/s11606-010-1526-3.

Hawn, C. 2009. Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. *Health Affairs* 28(2):361–368.

Hong, L. 2011. Tweets from justin bieber's heart: the dynamics of the "location" field in user profiles. In *CHI*.

Jain, S. H. 2009. Practicing medicine in the age of facebook. *New England Journal of Medicine* 361(7):649–651.

Lampos, V., and Cristianini, N. 2010. Tracking the flu pandemic by monitoring the social web. In *IAPR 2nd Workshop on Cognitive Information Processing (CIP 2010)*.

Lerman, K., and Ghosh, R. 2010. Information contagion: an empirical study of the spread of news on digg and twitter social networks. *CoRR* abs/1003.2664.

Minka, T. 2003. Estimating a Dirichlet distribution.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM*.

Paul, M., and Dredze, M. 2011. A model for mining public health topics from twitter. Technical report, Johns Hopkins University.

Paul, M., and Girju, R. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*.

Pelat, C.; Turbelin, C.; Bar-Hen, A.; Flahault, A.; and Valleron, A.-J. 2009. More diseases tracked by using google trends. *Emerg Infect Dis* 15(8):1327–1328.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *NAACL*.

Quincey, E., and Kostkova, P. 2010. Early warning and outbreak detection using social networking websites: The potential of twitter. In *Electronic Healthcare*. Springer Berlin Heidelberg.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *NAACL*.

Ritterman, J.; Osborne, M.; and Klein, E. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In *Workshop on Mining Social Media*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*.

Scanfeld, D.; Scanfeld, V.; and Larson, E. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control* 38(3):182–188.

Seifter, A.; Schwarzwalder, A.; Geis, K.; and Aucott, J. 2010. The utility of "google trends" for epidemiological research: Lyme disease as an example. *Geospatial Health* 4(2):135–137.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

Vance, K.; Howe, W.; and Dellavalle, R. P. 2009. Social internet sites as a source of public health information. *Dermatologic Clinics* 27(2):133–136.