A Model for Mining Public Health Topics from Twitter

Michael J. Paul and **Mark Dredze**
Human Language Technology Center of Excellence
Department of Computer Science
Johns Hopkins University
{mpaul,mdredze}@cs.jhu.edu

Abstract

We present the Ailment Topic Aspect Model (ATAM), a new topic model for Twitter that associates symptoms, treatments and general words with diseases (ailments). We train ATAM on a new collection of 1.6 million tweets discussing numerous health related topics. ATAM isolates more coherent ailments, such as influenza, infections, obesity, as compared to standard topic models. Furthermore, ATAM matches influenza tracking results produced by Google Flu Trends and previous influenza specialized Twitter models compared with government public health data.

1 Twitter and Public Health

Public health researchers dedicate considerable resources to population surveillance, which requires clinical encounters with health professionals. We propose a low cost alternative source for tracking public health trends: Twitter. Several studies have considered using Twitter for tracking various trends, including news tracking (Lerman and Ghosh, 2010; Petrović et al., 2010), earthquake monitoring (Sakaki et al., 2010), sentiment (Barbosa and Feng, 2010), and political opinions (Tumasjan et al., 2010; O'Connor et al., 2010). Similarly, tweets mention health related topics, such as "i got fever 102.5 i got flu i got sore eyes my throat hurts taking tylenol". This tweet indicates that the user has an ailment (flu), the associated symptoms (fever, etc.) and treatments (tylenol). Health self-reporting across millions of users can provide extensive real time information about population health.

In this work, we introduce a new method for extracting general public health information from millions of health related tweets. Previous work in this area has focused specifically on influenza, evaluating influenza surveillance (Lampos and Cristianini, 2010; Culotta, 2010b), analyzing tweets from the H1N1 pandemic (Quincey and Kostkova, 2010), and combining prediction markets and Twitter to predict H1N1 (Ritterman et al., 2009). These results all arise from supervised models built for specific applications (e.g. monitoring the flu.)

We present a more general approach that discovers many different ailments and learns symptom and treatment associations from tweets. Our first contribution is to create a data set of 1.6 million health related tweets (beyond just influenza.) To create structured information from these data, we develop a new topic model that organizes health terms into ailments, including associated symptoms and treatments. Our model uses explicit knowledge of symptoms and treatments to separate out coherent ailment groups from more general topics. We show that our model 1) discovers a larger number of more coherent ailments than LDA, 2) produces more detailed ailment information (symptoms/treatments) and 3) tracks disease rates consistent with published government statistics (influenza surveillance) despite the lack of supervised influenza training data.

2 A Twitter Health Corpus

We start with a collection of over 2 billion tweets from May 2009 to October 2010 (O'Connor et al., 2010). We first identify which of these messages contain health information. A first high recall keyword filter used a list of 20,000 keyphrases related to illnesses/diseases, symptoms, and treatments. [1] We removed retweets (marked with the "RT" tag) and tweets containing URLs; they were almost always false positives(e.g., news articles about the flu, rather than messages about a user's health.) The resulting set contained 11.7 million tweets.

Keyword filtering is insufficient since health keywords can be used in many contexts, e.g., "I'm *sick* of this" and "justin beber ur so cool and i have beber *fever*" (Culotta, 2010b). Instead, we obtain training data for a supervised classifier using Mechanical Turk (MTurk) (Callison-Burch and Dredze, 2010).We created a 5,128 tweet corpus labeled as *related* or *unrelated* to health. Turkers labeled tweets as:

- SICK: the message indicated that the user was sick with an acute illness (e.g., cold, flu).
- HEALTH: the message made general comments about the user or someone else's health (e.g., chronic conditions, lifestyle, diet).

Turkers also labeled messages as being UNRELATED to health, NOT ENGLISH, or AMBIGUOUS. Messages that were not about a particular person's health (e.g., news updates about the swine flu, advertisements for diet pills) were labeled as unrelated. We inserted gold labeled examples into the HITs to ensure that each message received 3 annotations from high quality turkers. The final label was determined by majority vote, removing examples where the majority of annotators were unsure of the best label. Our annotations differ from previous efforts which included only flu related tweets (Culotta, 2010b). The distribution of the labeled data is shown in Table 1.

We trained a binary SVM using SVM^{light} (Joachims, 1999) with a linear kernel and uni-gram, bi-gram, and tri-gram word features. SICK and HEALTH labeled messages were positive (36.1% of the examples), and UNRELATED and NOT-ENGLISH (63.9% of the examples) were negative. Tokenization treated contiguous blocks of punctuation as separators, and it helped to include these punctuation blocks as word tokens, rather than stripping them out. Hashtags and usernames were removed.

[1] Lists were scraped from `wrongdiagnosis.com/lists/{symptoms,condsaz,treats}.htm` and `mtworld.com/tools_resources/commondrugs.php`. We added two important keywords (*sick* and *doctor*) and removed some spurious keywords (e.g., *hat*).

Label	**2/3**	**3/3**
# tweets	5128	2753
SICK	24.4%	8.5%
HEALTH	11.3%	16.6%
UNRELATED	52.6%	58.1%
NOT ENGLISH	10.5%	16.7%
AMBIGUOUS	1.1%	0.1%

Table 1: The distribution of labels in our annotated data. The first column refers to tweets where the label is agreed on by two of the three annotators, while the second column requires agreement by all three annotators.

We favor a tagger with high precision over recall, which still yielded a large set of messages. We tuned the SVM slack parameter and prediction threshold using 10-fold cross validation to obtain a classifier with 90.4% precision and 32.0% recall. Applying this classifier to the 11.7 million messages produced a corpus of 1.63 million health related tweets.

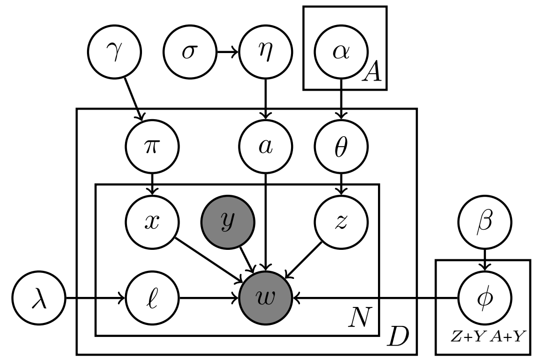
3 ATAM: A Model for Ailments in Twitter

We seek a model that can discover a range of health topics that are discussed in Twitter, not just a single disease. We turn to probabilistic topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003): generative models which associate word tokens with latent *topics* and discover latent structure in the data. Each document has a multinomial mixture over hidden topics, and each topic is defined by a multinomial distribution over words. Applying posterior inference over the model parameters given text typically yields topics where each topic's probability mass is assigned to words which frequently co-occur and have strong semantic relatedness.

Initial experiments with LDA produced some topics related to diseases, but most did not clearly indicate specific ailments. For example, many topics contained surgery terms, but it was not clear if these surgeries are associated with specific illnesses like physical injuries or cancer. In addition to topics, we require a model cognizant of the implicit structure of diseases. We develop a structured model that uses lists of symptoms and treatments to uncover diseases (ailments), such as flu, allergies, or cancer. An ailment contains both general words, as well as specific symptoms and treatments. Additionally, standard topics capture general terms unrelated to specific ailments.

For each health tweet, there is a latent ailment a drawn from a multinomial parameterized by η. Each ailment a indexes a distribution over words ϕ_a. Building on the idea behind the Topic Aspect Model (TAM) (Paul and Girju, 2010), a's distribution over words has three aspects, where an aspect y corresponds to a symptom (1), treatment (2) or a general word (0). The result is that each a indexes three multinomial distributions over words given by $\phi_{a,y}$ for $y \in \{0, 1, 2\}$. We assume symptoms and treatments are given by our web-scraped keyphrase lists described in section 2 (y is observed according to the presence of the phrase in our lists) since our goal is not to learn new symptoms or treatments – our lists are already extensive – but to learn symptom-treatment structures for ailments.[2] Our lists are phrases, while our model generates tokens. If a phrase appears in a list, we set the y value for each of the phrase's tokens, otherwise $y = 0$.

Not all word tokens fit into this symptom-treatment structure, because even when talking about ailments, users may use non-ailment vocabulary words (e.g. "home" and "watching" in the tweet "home with a fever watching TV.") Our model needs to account for "watching TV." Therefore, we create a set of Z topics, where each message contains a distribution θ over topics. As in LDA, these topic distributions are drawn from a Dirichlet distribution parameterized by α_a (each ailment has its own α vector to allow the model to make potential associations between the ailments and the various topics). To determine if a word is generated from an ailment dependent distribution $\phi_{a,y}$ or a non-ailment topic z, we include a switching variable $x \in \{0, 1\}$, sampled from a Binomial distribution parameterized by π, sampled on a per message basis from a Beta distribution parameterized by γ_0, γ_1. Having this set of non-ailment topics (i.e. topics that have only one multinomial of general words rather than three general/symptom/treatment aspects) is an addition over the original TAM structure. This idea is similar to



- Set the background switching binomial λ
- Draw an ailment distribution $\eta \sim$ Dir(σ)
- Draw word multinomials $\phi \sim$ Dir(β) for the topic, ailment, and background distributions
- For each message $1 \leq m \leq D$:
  - Draw a switching distribution $\pi \sim$ Beta(γ_0, γ_1)
  - Draw an ailment $a \sim$ Mult(η)
  - Draw a topic distribution $\theta \sim$ Dir(α_a)
  - For each word $w_i \in N_m$
    - Draw aspect $y_i \in \{0, 1, 2\}$ (observed)
    - Draw background switcher $\ell \in \{0, 1\} \sim$ Bi(λ)
    - If $\ell == 0$:
      - Draw $w_i \sim$ Mult($\phi_{B,y}$) (a background)
    - Else:
      - Draw $x_i \in \{0, 1\} \sim$ Bi(π)
      - If $x_i == 0$: (draw word from topic z)
        - Draw topic $z_i \sim$ Mult(θ)
        - Draw $w_i \sim$ Mult(ϕ_z)
      - Else: (draw word from ailment a aspect y)
        - Draw $w_i \sim$ Mult($\phi_{a,y}$)

Figure 1: The graphical model and generative process.

that of the Twitter conversation+topic model (Ritter et al., 2010) where words in a message can depend either on a LDA-style topic model or a message-specific conversation act, though the structure is different.

Finally, we include background distributions for common words, with a different distribution for each aspect. The switching variable ℓ (drawn from a Binomial parameterized by λ) determines if a word comes from the background or not. We call our model the **Ailment Topic Aspect Model** (ATAM, pronounced Atom). The graphical model and generative story are shown in Figure 1.

3.1 Inference

Parameter learning (posterior inference) can be carried out straightforwardly with Gibbs sampling. On

[2]Making y unobserved would force the model to discover treatments and symptoms, which we found unnecessary given the large available lists but could consider in future work.

each sampling iteration, new values of the random variables are sampled from a distribution conditioned on the current values of all other variables. As per the LDA framework, we can derive a collapsed Gibbs sampler by marginalizing the multinomials out of the sampling equations, requiring us to only sample the variables a, z, x and ℓ (Griffiths and Steyvers, 2004). We follow Ritter et al. (2010) and alternately sample the document-level variable a for a message and the token-level variables (z, x and ℓ). The sampling probabilities for z, x and ℓ are similar to the topic aspect model (Paul and Girju, 2010) and LDA, except that y is observed in our model and need not be sampled. We sample a according to:

$$\begin{aligned} &p(a_m|\mathbf{a_{-m}}, \mathbf{w}, \mathbf{y}, \mathbf{x}, \ell) \\ &\propto p(a_m|\mathbf{a_{-m}}) \prod_{n}^{N_m} p(w_{m,n}|\mathbf{a}, \mathbf{w}_{-(m,n)}, \mathbf{y}, \mathbf{x}, \ell) \end{aligned} \tag{1}$$

For each token in the message we sample the values for the variables from the distributions given by:

$$p(\ell = 0|a_m, \mathbf{w}, \mathbf{x}, \ell, \lambda, \beta) \propto \lambda p(w_i|\mathbf{w}_{-i}, \ell_{-i} = 0) \tag{2}$$

$$\begin{aligned} &p(\ell = 1, x = 0, z = k|a_m, \mathbf{w}, \mathbf{y}, \mathbf{x}, \ell, \alpha, \gamma, \lambda, \beta) \\ &\propto (1-\lambda)p(x_i = 0|\mathbf{w}_{-i}, \mathbf{x}_{-i}, \gamma)p(z_i = k|\mathbf{z}_{-i}, \alpha_{a_m}) \\ &p(w_i|z_i, \mathbf{w}_{-i}, \ell_{-i} = 1, \mathbf{x}_{-i} = 0, \beta) \end{aligned} \tag{3}$$

$$\begin{aligned} &p(\ell = 1, x = 1|a_m, \mathbf{w}, \mathbf{y}, \mathbf{x}, \ell, \gamma, \lambda, \beta) \\ &\propto (1-\lambda)p(x_i = 1|\mathbf{w}_{-i}, \mathbf{x}_{-i}, \gamma) \\ &p(w_i|\mathbf{a}, \mathbf{w}_{-i}, \mathbf{y}_{-i}\ell_{-i} = 1, \mathbf{x}_{-i} = 1, \beta) \end{aligned} \tag{4}$$

In these equations, the global index i denotes the token (m, n). We also need to infer the α_a vectors. To do this, we use a stochastic EM algorithm in which we update α after each sampling iteration based on the current variable assignments. We update this according to the fixed-point iteration algorithm given by Wallach (2006). Our model can easily be extended to a non-parametric model by using a Dirichlet Process prior for $p(a_m|\mathbf{a_{-m}})$.

Other hyperparameters were set manually: γ_0=1.0, γ_1=0.1 which gives a prior expectation that non-ailment topic words are more likely to appear in a tweet than ailment-specific words, and $\beta = 0.01$ and $\sigma = 100.0$ for the distributions over words and over ailments. The background switching Bernoulli is a fixed, user-defined distribution controlling the noise level – if we set this so that words have a high probability of coming from the background distribution, then they will only end up in a topic or ailment distribution if the pattern is very strong. Our initial observations suggest that tuning this parameter toward noise reduction is especially important for the symptom and treatment distributions for each ailment, where the most common words like "headache" and "surgery" will end up dominating every ailment, just as common words like "the" dominate a standard topic model without stop word removal or term weighting. We set $\lambda = 0.8$ (the probability that the word comes from the background). In all of our experiments, we ran for 8000 iterations with $Z = 15$ and $A = 20$.

4 Evaluation

In addition to ATAM, we trained an LDA model, augmented with a background topic (Chemudugunta et al., 2006) with the same probability as used by ATAM (0.8), for 5000 iterations with $Z = 35$ (the same number of clusters as our ATAM parameterization (Z+A)), and $\alpha = \beta = 0.01$. Three annotators each labeled the resulting LDA topics and ATAM ailments with either an ailment name or as "non-ailment" and we then obtained consensus as to the best label for each topic/ailment. Figure 3 shows 7 of the 15 ailments discovered by ATAM and labeled as ailments by annotators.[3]

We then evaluated model output through two MTurk experiments.[4] First, we measured agreement of turkers on labeling clusters (ailments/topics). We showed the top 8 words, 5 ailments and 5 treatments for a cluster.[5] We then showed three randomly sorted ailment names (one correct and two randomly chosen) as well as a "other" and "junk" option. 80 turkers provided annotations. ATAM discovered more ailments as measured by the number of ailments agreed to by two thirds of the annotators; 14 unique ATAM ailments versus 10 for LDA.

[3]The other eight ailment clusters were: upset stomach, flu, common cold, emergencies, infections, surgery, cancer, and skin problems.

[4]Experiments included gold HITs for quality control.

[5]We obtained symptoms/treatments for LDA topics by pulling out words that appeared in our keyphrase lists.

	Allergies	Insomnia	Obesity	Injuries	Respiratory	Dental	Aches/Pains
General	allergies	sleep	blood	knee	throat	ow	body
	nose	asleep	weight	leg	stop	teeth	need
	eyes	fell	eat	right	better	tooth	neck
	allergy	awake	healthy	ankle	voice	wisdom	hurts
	allergic	hours	fat	shoulder	hurts	dentist	head
Symptoms	sneezing	insomnia	pressure	pain	cough	pain	aches
	coughing	fall	weight	sore	coughing	toothache	pain
	cold	burning	loss	arthritis	cold	sore	sore
	nose	pain	blood	limping	sneezing	infection	muscle
	runny	falling	high	neck	sneeze	tooth	aching
Treatments	medicine	sleeping	diet	surgery	medicine	braces	massage
	benadryl	pills	exercise	brace	antibiotics	pain	exercise
	claritin	caffeine	dieting	crutches	codeine	relief	massages
	zyrtec	tylenol	insulin	physical	vitamin	muscle	bath
	drops	pill	exercising	therapy	tylenol	surgery	hot

Figure 2: Example output of the most likely words for ailments from the Ailment Topic Aspect Model. Ailment titles result from manual annotation of model output.

Still really sick. ER on Friday for **asthma** and <u>bronchitis</u>, strong *antibiotics* over wknd- back to Dr today. No work for the weary today!
came home early today, doctor says i have <u>strep throat</u> :(

Figure 3: Health related tweets labeled by ATAM as relating to the infections ailment. Underlined words are symptoms, italics are treatments, and bold are general ailment words. We also color code some words generated by the same non-ailment topics, where grey is the background topic.

Additionally, ATAM produced more identifiable ailments; 45% of turkers agreed with our gold LDA labels versus 70% for ATAM.

We next sought to evaluate which model produced more coherent ailment clusters. Using our labels, we paired ATAM and LDA clusters that represented the same ailment (e.g., both were labeled as flu.) We then displayed each ailment as before, but now side by side (randomly permuting which appeared on which side) with the ailment name (e.g. flu). 67 turkers were asked to select the list of words (including symptoms/treatments) that best described the given ailment, or to indicate a tie otherwise. ATAM was favored over LDA in 11 out of 18 comparisons with an average of 55% of the votes (median 64%). These experiments show that ATAM finds more unique ailments with higher coherence.

4.1 Syndromic Surveillance

One mission of the US Centers for Disease Control and Prevention (CDC) is syndromic surveillance: tracking disease rates in the population. The CDC publishes weekly influenza statistics under FluView.[6] While these statistics take considerable resources to produce, recent work has demonstrated that Web data and Twitter may be effective alternatives. 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 CDC's FluView (Carneiro and Mylonakis, 2009). Similar results have been reported for several other types of query logs (Valdivia et al., 2010; Polgreen et al., 2008; Hulth et al., 2009; Johnson et al., 2004; Pelat et al., 2009). Lampos and Cristianini (2010) are able to learn a Twitter flu rate producing a 0.97 correlation with the UK's Health Protection Agency influenza infection rates for the second half of 2009. Culotta (2010a) performed a similar analysis against CDC data (correlation 0.95 for September 2009 to May 2010). While these approaches explicitly model flu rates using regression against government

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

data, ATAM discovers ailments, including the flu. Can our general approach discover ailments that correlate with government data?

We computed the number of tweets per week assigned to ATAM's flu ailment, normalizing by the total number of tweets in the entire corpus for that week (August 2009 to May 2010).[7] The Pearson correlation coefficient between the flu frequencies in our data and the CDC data was 0.934, close to the previously reported methods specializing in influenza. For comparison, Google's Flu Trends (Google, 2011) computed from search query logs for the same time period yielded a nearly identical correlation of 0.932 with the CDC data.

5 Conclusion

We have demonstrated that public health information can be extracted from Twitter. We created a corpus of 5,128 messages labeled for relevance to health and produced a high precision labeling of 1.63 million English messages. Our Ailment Topic Aspect Model learns to group symptoms and treatments into latent ailments, as well as grouping remaining words into health related topics. We have shown that our model discovers meaningful ailment descriptions as well as ailments that can be used for syndromic surveillance.

We plan to incorporate richer models of Twitter tailored to specific problems, for example, by including temporal and geospatial dynamics to track diseases across a population. Public health information can be correlated with user location (Eisenstein et al., 2010), age and gender (Rao et al., 2010). Furthermore, by making the symptom/treatment aspect variable y partially observed, we could learn new symptoms or treatments, which may be of particular interest, such as if users have developed new home remedies. This could be coupled with a model for tagging diseases and symptoms.

[7] Earlier time periods contained gaps in the original data.

References

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING*.

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

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Workshop on Creating Speech and Language Data With Mechanical Turk at NAACL-HLT*.

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

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248.

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

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

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Empirical Natural Language Processing Conference (EMNLP)*.

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

Google. 2011. Google flu trends. `www.google.org/flutrends`.

T. Griffiths and M. Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4(2):e4378, 02.

Thorsten Joachims, 1999. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA.

HA Johnson, MM Wagner, WR Hogan, W Chapman, RT Olszewski, J Dowling, and G Barnas. 2004. Analysis of web access logs for surveillance of influenza. *Stud Health Technol Inform*, 107(Pt 2):1202–6.

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

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

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM*.

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

Camille Pelat, Clément Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. 2009. More diseases tracked by using google trends. *Emerg Infect Dis*, 15(8):1327–1328, August.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *NAACL*.

PM Polgreen, Y Chen, DM Pennock, and FD Nelson. 2008. Using internet searches for influenza surveillance. *Clin Infect Dis.*, 47(11):1443–8, December.

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

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Workshop on Search and Mining User-generated Contents (SMUC)*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *NAACL*.

Joshua Ritterman, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In *Workshop on Mining Social Media*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, New York, NY, USA.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

A Valdivia, J Lopez-Alcalde, M Vicente, M Pichiule, M Ruiz, and M Ordobas. 2010. Monitoring influenza activity in europe with google flu trends: comparison with the findings of sentinel physician networks - results for 2009-10. *Euro Surveill*, 15(29):19621, July.

H.M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *International Conference on Machine Learning*.