Separating Fact from Fear: Tracking Flu Infections on Twitter

Alex Lamb, Michael J. Paul, Mark Dredze
Human Language Technology Center of Excellence
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
{alamb3,mpaul19,mdredze}@jhu.edu

Abstract

Twitter has been shown to be a fast and reliable method for disease surveillance of common illnesses like influenza. However, previous work has relied on simple content analysis, which conflates flu tweets that report infection with those that express concerned awareness of the flu. By discriminating these categories, as well as tweets about the authors versus about others, we demonstrate significant improvements on influenza surveillance using Twitter.

1 Introduction

Twitter is a fantastic data resource for many tasks: measuring political (O'Connor et al., 2010; Tumasjan et al., 2010), and general sentiment (Bollen et al., 2011), studying linguistic variation (Eisenstein et al., 2010) and detecting earthquakes (Sakaki et al., 2010). Similarly, Twitter has proven useful for public health applications (Dredze, 2012), primarily disease surveillance (Collier, 2012; Signorini et al., 2011), whereby public health officials track infection rates of common diseases. Standard government data sources take weeks while Twitter provides an immediate population measure.

Strategies for Twitter influenza surveillance include supervised classification (Culotta, 2010b; Culotta, 2010a; Eiji Aramaki and Morita, 2011), unsupervised models for disease discovery (Paul and Dredze, 2011), keyword counting[1], tracking geographic illness propagation (Sadilek et al., 2012b), and combining tweet contents with the social network (Sadilek et al., 2012a) and location information (Asta and Shalizi, 2012). All of these methods rely on a relatively simple NLP approach to analyzing the tweet content, i.e. n-gram models for classifying related or not related to the flu. Yet examining flu tweets yields a more complex picture:

- going over to a friends house to check on her son. he has the flu and i am worried about him
- Starting to get worried about swine flu...

Both are related to the flu and express worry, but tell a different story. The first reports an infection of another person, while the second expresses the author's concerned awareness. While infection tweets indicate a rise in infection rate, awareness tweets may not. Automatically making these distinctions may improve influenza surveillance, yet requires more than keywords.

We present an approach for differentiating between flu infection and concerned awareness tweets, as well as self vs other, by relying on a deeper analysis of the tweet. We present our features and demonstrate improvements in influenza surveillance.

1.1 Related Work

Much of the early work on web-based influenza surveillance relied on query logs and click-through data from search engines (Eysenbach, 2006), most famously Google's Flu Trends service (Ginsberg et al., 2008; Cook et al., 2011). Other sources of information include articles from the news media and online mailing lists (Brownstein et al., 2010).

2 Capturing Nuanced Trends

Previous work has classified messages as being related or not related to influenza, with promising surveillance results, but has ignored nuanced differences between flu tweets. Tweets that are related to

[1]The DHHS competition relied solely on keyword counting. http://www.nowtrendingchallenge.com/

flu but do not report an infection can corrupt infection tracking.

Concerned Awareness vs. Infection (A/I) Many flu tweets express a concerned awareness as opposed to infection, including fear of getting the flu, an awareness of increased infections, beliefs related to flu infection, and preventative flu measures (e.g. flu shots.) Critically, these people do not seem to have the flu, whereas infection tweets report having the flu. This distinction is similar to modality (Prabhakaran et al., 2012a). Conflating these tweets can hurt surveillance, as around half of our annotated flu messages were awareness. Identifying awareness tweets may be of use in-and-of itself, such as for characterizing fear of illness (Epstein et al., 2008; Epstein, 2009), public perception, and discerning sentiment (e.g. flu is negative, flu shots may be positive.) We focus on surveillance improvements.[2]

Self vs. Other (S/O) Tweets for both awareness and infection can describe the author (self) or others. It may be that self infection reporting is more informative. We test this hypothesis by classifying tweets as self vs. other.

Finding Flu Related Tweets (R/U) We must first identify messages that are flu related. We construct a classifier for flu related vs. unrelated.

3 Features

Token sequences (n-grams) are an insufficient feature set, since our classes share common vocabularies. Consider,

- A little worried about the swine flu epidemic!
- Robbie might have swine flu. I'm worried.

Both tweets mention flu and worried, which distinguish them as flu related but not specifically awareness or infection, nor self or other. Motivated by Bergsma et al. (2012), we complement 3-grams with additional features that capture longer spans of text and generalize using part of speech tags. We begin by processing each tweet using the ARK POS tagger (Gimpel et al., 2011) and find phrase segmentations using punctuation tags.[3] Most phrases were two (31.2%) or three (26.6%) tokens long.

[2] While tweets can both show awareness and report an infection, we formulate a binary task for simplicity since only a small percentage of tweets were so labeled.

[3] We used whitespace for tokenization, which did about the same as Jerboa (Van Durme, 2012).

Class Name	Words in Class
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Possession	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
Past Tense	was, did, had, got, were, or verb with the suffix "ed"
Present Tense	is, am, are, have, has, or verb with the suffix "ing"
Self	I, I've, I'd, I'm, im, my
Others	your, everyone, you, it, its, u, her, he, she, he's, she's, she, they, you're, she'll, he'll, husband, wife, brother, sister, your, people, kid, kids, children, son, daughter

Table 1: Our manually created set of word class features.

Word Classes For our task, many word types can behave similarly with regard to the label. We create word lists for possessive words, flu related words, fear related words, "self" words, "other" words, and fear words (Table 1). A word's presence triggers a count-based feature corresponding to each list.

Stylometry We include Twitter-specific style features. A feature is included for retweet, hashtags, and mentions of other users. We include a feature for emoticons (based on the emoticon part-of-speech tag). We include a more specific feature for positive emoticons (:) :D :)). We also include a feature for negative emoticons (:(:/). Additionally, we include a feature for links to URLs.

Part of Speech Templates We include features based on a number of templates matching specific sequences of words, word classes, and part of speech tags. Where any word included in the template matches a word in one of the word classes, an additional feature is included indicating that the word class was included in that template.

- Tuples of (subject,verb,object) and pairs of (subject, verb), (subject, object), and (verb, object). We use a simple rule to construct these tuples: the first noun or pronoun is taken as the subject, and the first verb appearing after the subject is taken as the verb. The object is taken as any noun or pronoun that appears before a verb or at the end of a phrase.

• A pairing of the first pronoun with last noun. These are useful for S/O, e.g. *I am worried that my son has the flu* to recognize the difference between the author (I) and someone else.
• Phrases that begin with a verb (pro-drop). This is helpful for S/O, e.g. *getting the flu!* which can indicate self even without a self-related pronoun. An additional feature is included if this verb is past-tense.
• Numeric references. These often indicate awareness (number of people with the flu) and are generally not detected by an n-gram model. We add a separate feature if the word following has the root "died", e.g. *So many people dying from the flu, I'm scared!*
• Pair of first pronoun/noun with last verb in a phrase. Many phrases have multiple verbs, but the last verb is critical, e.g. *I had feared the flu.* Additional features are added if the noun/pronoun is in the "self" or "other" word class, and if the verb is in the "possessive" word class.
• Flu appears as a noun before first verb in a phrase. This indicates when flu is a subject, which is more likely to be about awareness.
• Pair of verb and following noun. This indicates the verbs object, which can change the focus of A/I, e.g., *I am getting a serious case of the flu* vs. *I am getting a flu shot.* Additional features are added if the verb is past tense (based on word list and suffix "-ed".)
• Whether a flu related word appears as a noun or an adjective. When flu is used as an adjective, it may indicate a more general discussion of the flu, as opposed to an actual infection *I hate this flu* vs. *I hate this flu hype.*
• If a proper noun is followed by a possessive verb. This may indicate others for the S/O task *Looks like Denmark has the flu.* An additional feature fires for any verb that follows a proper noun and any past tense verb that follows a proper noun.
• Pair each noun with "?". While infection tweets are often statements and awareness questions, the subject matters, e.g. *Do you think that swine flu is coming to America?* as awareness. An equivalent feature is included for phrases ending with "!".

While many of our features can be extracted using a syntactic parser (Foster et al., 2011), tweets are very short, so our simple rules and over-generating features captures the desired effects without parsing.

	Self	Other	Total
Awareness	23.15%	24.07%	47.22%
Infection	37.21%	15.57%	52.78%
Total	60.36%	39.64%	

Table 2: The distribution over labels of the data set. Infection tweets are more likely to be about the author (self) than those expressing awareness.

3.1 Learning

We used a log-linear model from Mallet (McCallum, 2002) with L_2 regularization. For each task, we first labeled tweets as related/not-related and then classified the related tweets as awareness/infection and self/others. We found this two phase approach worked better than multi-class.

4 Data Collection

We used two Twitter data sets: a collection of 2 billion tweets from May 2009 and October 2010 (O'Connor et al., 2010)[4] and 1.8 billion tweets collected from August 2011 to November 2012. To obtain labeled data, we first filtered the data sets for messages containing words related to concern and influenza,[5] and used Amazon Mechanical Turk (Callison-Burch and Dredze, 2010) to label tweets as concerned awareness, infection, media and unrelated. We allowed multiple categories per tweet. Annotators also labeled awareness/infection tweets as self, other or both. We included tweets we annotated to measure Turker quality and obtained three annotations per tweet. More details can be found in Lamb et al. (2012).

To construct a labeled data set we removed low quality annotators (below 80% accuracy on gold tweets.) This seemed like a difficult task for annotators as a fifth of the data had no annotations after this step. We used the majority label as truth and ties were broken using the remaining low quality annotators. We then hand-corrected all tweets, changing 13.5% of the labels. The resulting data set contained 11,990 tweets (Table 2), 5,990 from 2011-2012 for training and the remaining from 2009-2010 as test.[6]

[4]This coincided with the second and larger H1N1 (swine flu) outbreak of 2009; swine flu is mentioned in 39.6% of the annotated awareness or infection tweets.

[5]e.g. "flu", "worried", "worry", "scared", "scare", etc.

[6]All development was done using cross-validation on training data, reserving test data for the final experiments.

Feature Removed	A/I	S/O
n-grams	0.6701	0.8440
Word Classes	**0.7735**	0.8549
Stylometry	0.8011	**0.8522**
Pronoun/Last Noun	0.7976	0.8534
Pro-Drop	0.7989	0.8523
Numeric Reference	0.7988	0.8530
Pronoun/Verb	0.7987	0.8530
Flu Noun Before Verb	0.7987	0.8526
Noun in Question	0.8004	0.8534
Subject,Object,Verb	0.8005	0.8541

Table 3: F1 scores after feature ablation.

5 Experiments

We begin by evaluating the accuracy on the binary classification tasks and then measure the results from the classifiers for influenza surveillance. We created precision recall curves on the test data (Figure 1), and measured the highest F1, for the three binary classifiers. For `A/I` and `S/O`, our additional features improved over the n-gram baselines. We performed feature ablation experiments (Table 3) and found that for `A/I`, the word class features helped the most by a large margin, while for `S/O` the stylometry and pro-drop features were the most important after n-grams. Interestingly, `S/O` does equally well removing just n-gram features, suggesting that the `S/O` task depends on a few words captured by our features.

Since live data will have classifiers run in stages – to filter out not-related tweets – we evaluated the performance of two-staged classification. F1 dropped to 0.7250 for `A/I` and `S/O` dropped to 0.8028.

5.1 Influenza surveillance using Twitter

We demonstrate how our classifiers can improve influenza surveillance using Twitter. Our hypothesis is that by isolating infection tweets we can improve correlations against government influenza data. We include several baseline methods:

Google Flu Trends: Trends from search queries.[7]

Keywords: Tweets that contained keywords from the DHHS Twitter surveillance competition.

ATAM: We obtained 1.6 million tweets that were automatically labeled as influenza/other by ATAM (Paul and Dredze, 2011). We trained a binary classifier with n-grams and marked tweets as flu infection.

[7] http://www.google.org/flutrends/

Data	System	2009	2011
Google	Flu Trends	0.9929	0.8829
Twitter	ATAM	0.9698	0.5131
	Keywords	0.9771	0.6597
	All Flu	0.9833	0.7247
	Infection	**0.9897**	**0.7987**
	Infection+Self	0.9752	0.6662

Table 4: Correlations against CDC ILI data: Aug 2009-Aug 2010, Dec 2011 to Aug 2012.

We evaluated three trends using our three binary classifiers trained with a reduced feature set close to the n-gram features:[8]

All Flu: Tweets marked as flu by Keywords or ATAM were then classified as related/unrelated.[9] This trend used all flu-related tweets.

Infection: Related tweets were classified as either awareness or infection. This used infection tweets.

Infection+Self: Infection were then labeled as self or other. This trend used self tweets.

All five of these trends were correlated against data from the Centers for Disease Control and Prevention (CDC) weekly estimates of influenza-like illness (ILI) in the U.S., with Pearson correlations computed separately for 2009 and 2011 (Table 4).[10] Previous work has shown high correlations for 2009 data, but since swine flu had so dominated social media, we expect weaker correlations for 2011.

Results are show in Table 4 and Figure 2 shows two classifiers against the CDC ILI data. We see that in 2009 the Infection curve fits the CDC curve very closely, while the All Flu curve appears to substantially overestimate the flu rate at the peak. While 2009 is clearly easier, and all trends have similar correlations, our Infection classifier beats the other Twitter methods. All trends do much worse in

[8] Classifiers trained on 2011 data and thresholds selected to maximize F1 on held out 2009 data.

[9] Since our data set to train related or unrelated focused on tweets that appeared to mention the flu, we first filtered out obvious non-flu tweets by running ATAM and Keywords.

[10] While the 2009 data is a 10% sample of Twitter, we used a different approach for 2011. To increase the amount of data, we collected Tweets mentioning health keywords and then normalized by the public stream counts. For our analysis, we excluded days that were missing data. Additionally, we used a geolocator based on user provided locations to exclude non-US messages. See (Dredze et al., 2013) for details and code for the geolocator.

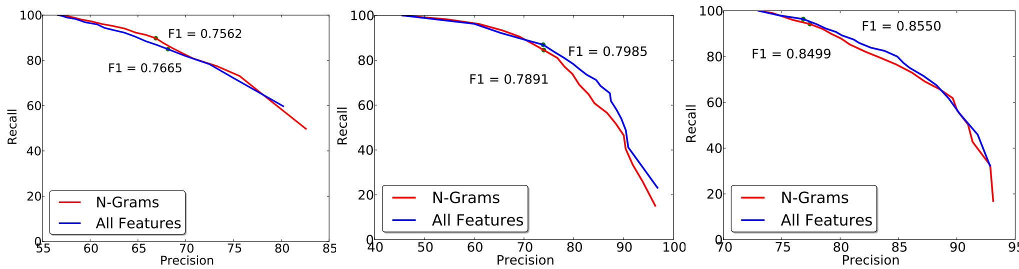

Figure 1: Left to right: Precision-recall curves for related vs. not related, awareness vs. infection and self vs. others.

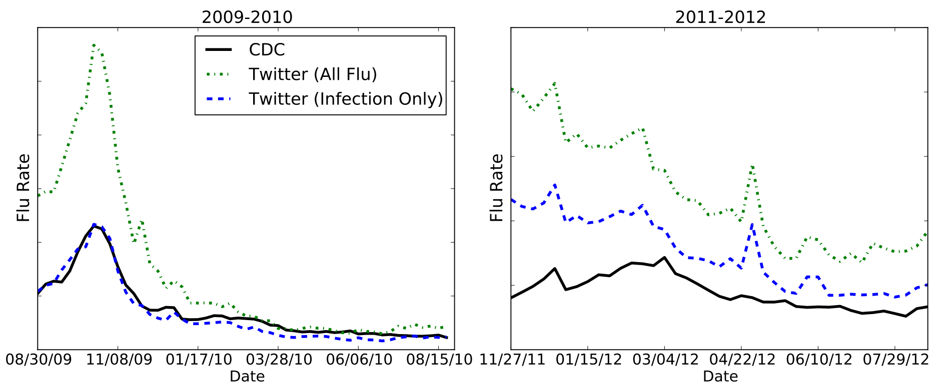

Figure 2: The Twitter flu rate for two years alongside the ILI rates provided by the CDC. The y-axes are not comparable between the two years due to differences in data collection, but we note that the 2011-12 season was much milder.

the 2011 season, which was much milder and thus harder to detect. Of the Twitter methods, those using our system were dramatically higher, with the Infection curve doing the best by a significant margin. Separating out infection from awareness (`A/I`) led to significant improvements, while the `S/O` classifier did not, for unknown reasons.

The best result using Twitter reported to date has been by Doan et al. (2012), whose best system had a correlation of 0.9846 during the weeks beginning 8/30/09–05/02/10. Our Infection system had a correlation of 0.9887 during the same period. While Google does better than any of the Twitter systems, we note that Google has access to much more (proprietary) data, and their system is trained to predict CDC trends, whereas our Twitter system is intrinsically trained only on the tweets themselves.

Finally, we are also interested in daily trends in addition to weekly, but there is no available evaluation data on this scale. Instead, we computed the stability of each curve, by measuring the day-to-day changes. In the 2009 season, the relative increase or decrease from the previous day had a variance of 3.0% under the Infection curve, compared to 4.1% under ATAM and 6.7% under Keywords.

6 Discussion

Previous papers have implicitly assumed that flu-related tweets mimick the infection rate. While this was plausible on 2009 data that focused on the swine flu epidemic, it is clearly false for more typical flu seasons. Our results show that by differentiating between types of flu tweets to isolate reports of infection, we can recover reasonable surveillance. This result delivers a promising message for the NLP community: deeper content analysis of tweets matters. We believe this conclusion is applicable to numerous Twitter trend tasks, and we encourage others to investigate richer content analyses for these tasks. In particular, the community interested in modeling author beliefs and influence (Diab et al., 2009; Prabhakaran et al., 2012b; Biran and Rambow, 2011) may find our task and data of interest. Finally, beyond surveillance, our methods can be used to study disease awareness and sentiment, which has implications for how public health officials respond to outbreaks. We conclude with an example of this distinction. On June 11th, 2009, the World Health Organization declared that the swine flu had become a global flu pandemic. On that day, flu awareness increased 282%, while infections increased only 53%.

References

Dena Asta and Cosma Shalizi. 2012. Identifying influenza trends via Twitter. In *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models and Applications*.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proc. NAACL-HLT*, pages 327–337.

O. Biran and O. Rambow. 2011. Identifying justifications in written dialogs. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 162–168. IEEE.

J. Bollen, A. Pepe, and H. Mao. 2011. Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453.

John S. Brownstein, Clark C. Freifeld, Emily H. Chan, Mikaela Keller, Amy L. Sonricker, Sumiko R. Mekaru, and David L. Buckeridge. 2010. Information technology and global surveillance of cases of 2009 h1n1 influenza. *New England Journal of Medicine*, 362(18):1731–1735.

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

N. Collier. 2012. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, 7(7):731–749.

Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. 2011. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLOS ONE*, 6(8):e23610.

A. Culotta. 2010a. Towards detecting influenza epidemics by analyzing Twitter messages. In *ACM Workshop on Soc.Med. Analytics*.

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

Mona T. Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *ACL Third Linguistic Annotation Workshop*.

S. Doan, L. Ohno-Machado, and N. Collier. 2012. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. *arXiv preprint arXiv:1210.0848*.

Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. A Twitter geolocation system with applications to public health. Working paper.

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

Sachiko Maskawa Eiji Aramaki and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Empirical Natural Language Processing Conference (EMNLP)*.

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)*.

Joshua Epstein, Jon Parker, Derek Cummings, and Ross Hammond. 2008. Coupled contagion dynamics of fear and disease: Mathematical and computational explorations. *PLoS ONE*, 3(12).

J.M. Epstein. 2009. Modelling to contain pandemics. *Nature*, 460(7256):687–687.

G. Eysenbach. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium*, pages 244–248. AMIA.

J. Foster, Ö. Çetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan, J. Van Genabith, et al. 2011. # hardtoparse: Pos tagging and parsing the Twitterverse. In *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Association for Computational Linguistics (ACL)*.

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.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2012. Investigating Twitter as a source for studying behavioral responses to epidemics. In *AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*.

A.K. McCallum. 2002. MALLET: A machine learning for language toolkit.

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

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012a.

Statistical modality tagging from rule-based annotations and crowdsourcing. In *Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM 2012)*.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012b. Predicting overt display of power in written dialogs. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012a. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012b. Predicting disease transmission from geotagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

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.

A. Signorini, A.M. Segre, and P.M. Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic. *PLoS One*, 6(5):e19467.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media*, pages 178–185.

B. Van Durme. 2012. Jerboa: A toolkit for randomized and streaming algorithms. Technical report, Technical Report 7, Human Language Technology Center of Excellence, Johns Hopkins University.