Worldwide Influenza Surveillance through Twitter

Michael J. Paul[a], **Mark Dredze**[a], **David A. Broniatowski**[b], **Nicholas Generous**[c]
[a] Human Language Technology Center of Excellence, Johns Hopkins University; Baltimore, MD 21218
[b] Department of Engineering Management and Systems Engineering, George Washington University; Washington, DC 20052
[c] Defense Systems and Analysis Division, Los Alamos National Laboratory; Los Alamos, NM 87545

Abstract

We evaluate the performance of Twitter-based influenza surveillance in ten English-speaking countries across four continents. We find that tweets are positively correlated with existing surveillance data provided by government agencies in these countries, with r values ranging from .37–.81. We show that incorporating Twitter data into a strong autoregressive baseline reduces mean squared error in 80 to 100 percent of locations depending on the lag, with larger improvements when reporting delays are longer.

Introduction

Web-based data streams, including search engine statistics and social media messages, have emerged as complementary sources of data for influenza surveillance. Tweets – status updates from the microblog Twitter – are a particularly promising data source due to the large volume and openness of the platform (Broniatowski, Paul, and Dredze 2014). Tweets can often be resolved to geographic locations (Oussalah et al. 2012), either through GPS coordinates (e.g. when used with mobile devices), self-reported locations in user profiles (Hecht et al. 2011), or content based geolocation (Wing and Baldridge 2011; Han, Cook, and Baldwin 2014). This property makes Twitter suitable for surveillance across multiple geographic locations.

Most research evaluating Twitter for influenza surveillance has focused on the United States – at the national (Culotta 2010; Chew and Eysenbach 2010; Signorini, Segre, and Polgreen 2011; Doan, Ohno-Machado, and Collier 2012) and local levels (Broniatowski, Paul, and Dredze 2013; Nagar et al. 2014). Other countries have been explored to a lesser extent, including the United Kingdom (Lampos and Cristianini 2012; Dredze et al. 2013), Japan (Eiji Aramaki and Morita 2011), Portugal (Santos and Matos 2014), and China (through microblog Sina Weibo) (Sun, Ye, and Ren 2014). All of these Twitter systems have been evaluated on just one or a small number of locations.

In this study, we evaluate the utility of Twitter data for surveillance in several countries around the globe. We use the influenza detection system of Lamb, Paul, and Dredze (2013), which has been shown to have state of the art performance at Twitter-based flu tracking in the US (Broniatowski, Paul, and Dredze 2013). Our new experiments show that this system is moderately to highly correlated with government-provided surveillance data in ten countries. We show that Twitter provides utility in influenza prediction compared to a strong autoregressive baseline.

Data

Twitter Influenza Surveillance

We use our Twitter influenza surveillance system described in Lamb, Paul, and Dredze (2013), and Broniatowski, Paul, and Dredze (2013). The detection algorithm categorizes individual tweets for relevance to influenza infection and then produces estimates by aggregating the relevant tweets over some time interval (e.g. weekly).

Tweets are categorized according to a cascade of three logistic regression models that classify tweets based on different granularities of relevance: whether a tweet is relevant to health, whether a health tweet is about influenza, and whether an influenza tweet indicates an infection as opposed to a general awareness of the flu. The first model is trained on 5,128 hand-labeled English-language tweets, while the two influenza models are trained on 11,990 tweets.[1] The classifiers use a variety of features, including n-grams, indicators of URLs and @-mentions, and a variety of shallow syntactic features created using part-of-speech tag templates (Gimpel et al. 2011).

Our Twitter dataset contains approximately 4 million public tweets per day starting November 27, 2011. More details of our Twitter data collection are described in Broniatowski et al. (2013). We create weekly estimates of influenza prevalence as the number of influenza infection tweets classified by our models, normalized by the total number of public tweets in the week (the *influenza rate* – the infection prevalence per tweet). The denominator adjusts the Twitter rate to account for changes in overall tweet volume over time.

Estimates for particular locations are produced by restricting the numerator and denominator only to tweets posted

[1] Because the models were trained on tweet data, rather than relying on pre-specified keywords, we can in principle handle mismatches in medical terminology between professionals and individual web users (Nie et al. 2014a; 2014b).

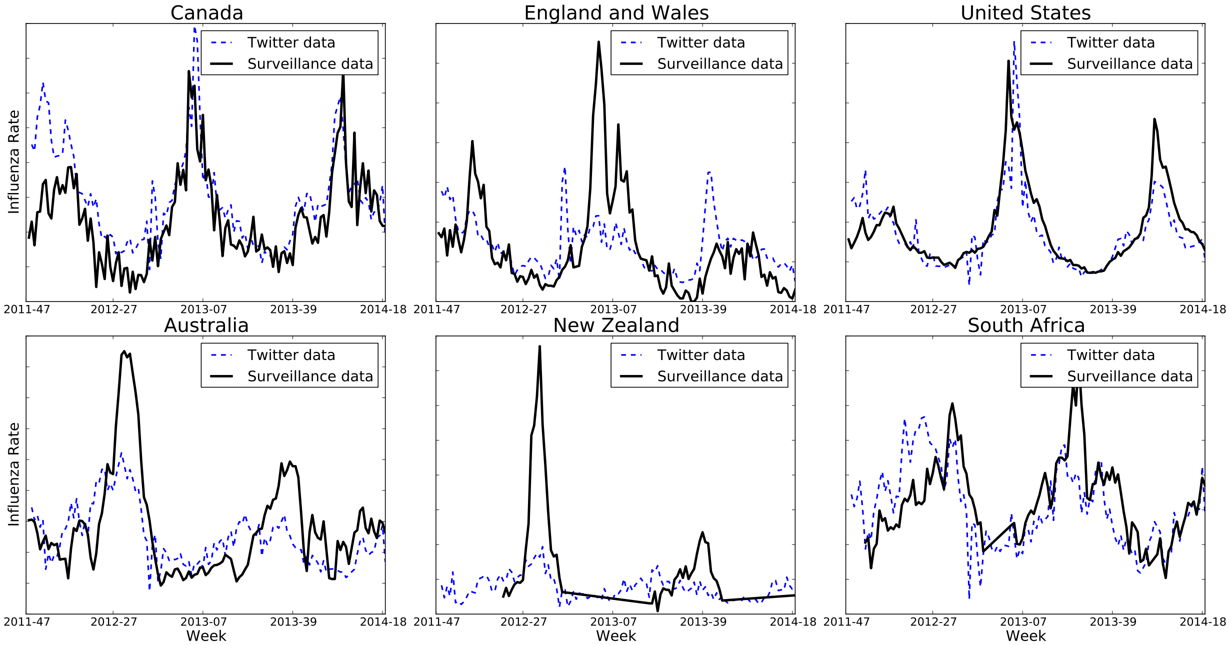

Figure 1: The influenza rate over time as measured by official government surveillance data and our Twitter surveillance estimates for three northern hemisphere countries (top) and three southern hemisphere countries (bottom).

from the specified location. Tweets are resolved to specific locations using the *Carmen* geolocation system (Dredze et al. 2013), which is estimated to resolve countries with 91% accuracy.

The data were downloaded from HealthTweets.org (Dredze et al. 2014), a public website that provides weekly estimates produced by our system.[2]

Sentinel Influenza Surveillance

We collected and compared to sentinel surveillance data publicly available from various government agencies. Our dataset includes national estimates of influenza-like illness (ILI) for ten countries: Australia, Canada, Ireland, New Zealand, South Africa, the United Kingdom (England and Wales, Wales alone, Scotland, Northern Ireland), and the United States (US).

We selected these countries because their populations are primarily English-speaking and there was available weekly or biweekly government data for influenza during the flu season.

The weekly data span three influenza seasons, beginning week 47 of 2011 (when our Twitter dataset begins) through week 18 of 2014, when the data were collected.[3] We have less data for some locations. New Zealand begins later and is particularly sparse because the agency does not release reports in off-season weeks. Other locations have missing data as well, either because of limited reporting off season, or anomalous gaps due to various circumstances. Our table of results therefore provides both the range of weeks as well as well as the number of data points that are included.

Different agencies release reports with varying delays, which we characterize in terms of a range of days. For example, in the US the Centers for Disease Control and Prevention (CDC) releases weekly reports every Friday for the previous week. Since the reporting weeks end on Saturday, we classify the delay as 6–13 days. The delay is thus 1–2 weeks. Most researchers have used the upper end and assumed a 2-week lag in their models (Ginsberg et al. 2008; Goel et al. 2010; Lazer et al. 2014), but both are realistic. Some agencies release reports with less frequency (e.g. biweekly), so the range of days is wider.

Overall, it is clear that this data set is an imperfect measure of influenza and of variable quality. Nevertheless, it is an accurate reflection of the current state of practice at a wide variety of health agencies. Improvements enabled by our Twitter system to these methods mean real improvements in surveillance for health officials and clinicians.

Full details of these data are provided in Table 3 at the end of this paper.

Model Evaluation

Following previous work (Culotta 2010; Broniatowski, Paul, and Dredze 2013), we report the Pearson correlation between the Twitter estimates and the values provided by government agencies.

Additionally, we recently noted (Paul, Dredze, and Broniatowski 2014b; 2014a) that simple autoregressive models – regression models that predict the current week value based

[2] The website requires an account, which can be freely requested using the form linked from the homepage.

[3] We have more recent data for some locations, but we restricted the time range to be consistent across all locations, so that the results are comparable.

Location	Pop.	# Tweets / week		Delay	n	Time Range	r	MSE Red. (%)		
		All	Flu					ℓ=1	ℓ=2	ℓ=3
Australia	22.7m	60,332	304	12–26	125	201149 – 201418	0.648*	10.5	19.5*	29.2*
Canada	34.9m	137,608	544	6–20	125	201149 – 201418	0.740*	7.7	24.1*	37.2*
England+Wales	56.1m	339,387	1,935	4–11	122	201149 – 201418	0.517*	-0.6	7.8	9.5
Ireland	4.6m	30,180	211	4–11	113	201149 – 201418	0.433*	1.6	5.4	7.4
New Zealand	4.4m	9,003	46	3–10	45	201220 – 201344	0.614*	18.0	37.0	59.2
Northern Ireland	1.8m	6,415	46	4–11	122	201149 – 201418	0.422*	5.3	6.1	8.7
Scotland	5.3m	32,212	184	4–11	122	201149 – 201418	0.517*	-3.2	-0.5	4.2
South Africa	51.2m	33,095	495	>30	105	201203 – 201418	0.547*	5.6	17.2	25.3
United States	314.0m	2.1m	5,846	6–13	125	201149 – 201418	0.814*	15.3	17.7	33.6*
Wales	3.1m	14,169	96	4–11	122	201149 – 201418	0.374*	2.7	6.8	3.3

Table 1: For each location, the table shows the Pearson correlation coefficient (r) between the surveillance data and the Twitter rates, as well as the relative reduction in mean squared error (MSE Red.) when incorporating Twitter data into the nowcasting model with lags (ℓ) of 1–3 weeks. Additionally, the table includes details about the locations, including the population (Pop.) and average weekly tweet volume (all tweets and flu tweets), the reporting delay (Delay) for that location, measured as a range of days, and the time span (given by year and week number) and number of data points (n) included.
* indicates significance with $p < 0.05$. Significance of error reduction is measured with a paired t-test of weekly error values.

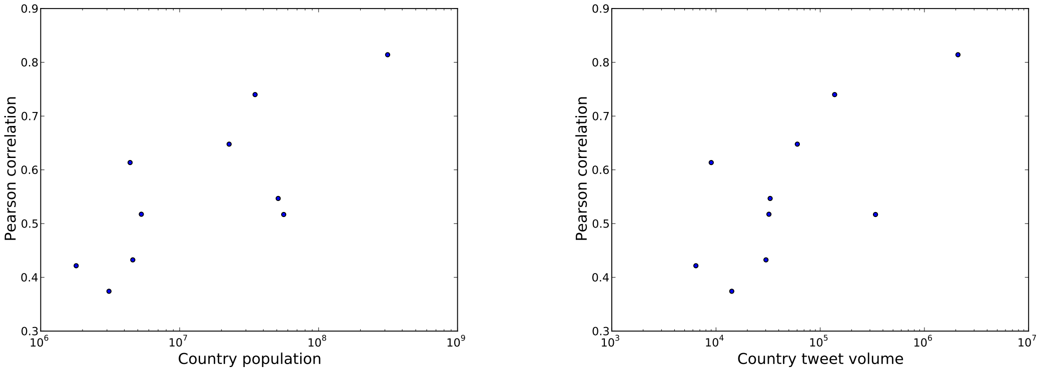

Figure 2: Each country's Pearson correlation (r) value as reported in Table 1 (y-axis) along with the population (left) or tweet volume (right) for each country (x-axis; log scale).

on previous weeks – are very strong baseline models, and in fact are better at nowcasting the current influenza rate better than Twitter alone. Even if the Twitter data is highly correlated, it does not necessarily add predictive power beyond the original time series. For this reason, we will compare to a baseline linear autoregressive model that estimates the value y_w of the current week w given the previous two weeks of available data:

$$\hat{y}_w = \alpha_0 y_{w-\ell} + \alpha_1 y_{w-\ell-1} \tag{1}$$

The variable ℓ denotes the lag in reporting, e.g. $\ell = 1$ if the surveillance data is released one week later. We will investigate performance with lags of 1–3 weeks. The α coefficients are estimated via standard least squares regression. This model can be used to "nowcast" the current week's value given the previous data (Lampos and Cristianini 2012; Achrekar et al. 2012).

We then experiment with a similar model that includes the Twitter estimate z_w for week w as a predictor:

$$\hat{y}_w = \gamma z_w + \alpha_0 y_{w-\ell} + \alpha_1 y_{w-\ell-1} \tag{2}$$

The parameters were estimated using Scikit-learn (Pedregosa et al. 2011). We used Ridge Regression with a tiny penalty on the ℓ_2 norm (10^{-6}) simply to stabilize the parameter values. Weeks with missing data were excluded from our analysis.

We will compare the predictive performance of these two nowcasting models using five-fold cross validation across the entire span of time. We segment the data into five contiguous sets, train the models on 80% of the data at a time, and evaluate on the remaining 20%, repeated across all folds. This experiment is repeated for each country. By evaluating on out-of-sample data, cross-validation provides better estimates of how the models will perform in the future, compared to measures only on in-sample data.

Results

Table 1 gives our key results, including the correlation between the Twitter rates and surveillance data, and the reduction in mean squared error (MSE) when incorporating Twitter into the nowcasting model.

The Twitter data is moderately to highly correlated with the surveillance data: a significant positive correlation exists for every location. The Twitter data reduces nowcasting error in nearly all cases: 8 out of 10 locations with ℓ=1, 9 out of 10 with ℓ=2, and all 10 with ℓ=3.

In addition to the reduction in MSE, we show the original MSE values (for both the baseline model and the model with Twitter) in Table 2. In addition to the raw MSE, we show a version of MSE that has been normalized by the data variance, so that these values can be compared across locations. Finally, we also measured mean relative error, following Lazer et al. (2014), shown in the rightmost columns of Table 2. The results are very different when measured with relative error rather than with MSE. Under this metric, Twitter reduces error in only 20% of locations. The United States has the largest (and only significant) reduction in relative error. This is also the location studied by Lazer et al., analyzing Google Flu Trends.

We investigated the relationship between the correlation value r in the table and the population or tweet volume of each country. Figure 2 shows that there is a log-linear relationship between the performance and the country population (a correlation of .764 between the r values and the log of the population) and similarly with overall tweet volume (correlation of .701). It thus seems that Twitter surveillance generally works better in locations with larger populations and higher Twitter activity, though there are exceptions. For instance, Twitter correlates surprisingly poorly with England+Wales for having the second-highest population and tweet volume, with r on the lower end of our results, at .517.

Discussion

Our results show that Twitter influenza data is well correlated with data from ten English-speaking countries across four continents.

That tweets are well correlated is important independent of the nowcasting performance. The nowcasting models provide a benchmark to demonstrate the marginal utility of adding Twitter to existing data, yet these models cannot always be used in practice, and therefore this baseline is perhaps unrealistically strong for many scenarios. For example, these models are not robust to gaps, and cannot be used during off-season weeks in locations that do not report during those periods. Our results ignore weeks that are missing from the surveillance data, yet in practice tweets could be useful when these data are missing.

The nowcasting experiments are also unrealistic because they assume that the data used for nowcasting were available at the time the prediction is made. In fact, the numbers that are initially published by agencies are often revised in the future, e.g. as additional providers in the network will submit reports. The effect of revisions is not trivial. Our recent retrospective study of US data (Paul, Dredze, and Broniatowski 2014b; 2014a) showed that nowcasting models which used the revised values (not available at the time of the nowcast) underestimated the error by 42% over using the values that were initially reported. Moreover, including Twitter data reduced nowcasting error by 30% when using the initial values, but only by 6% when inaccurately using the revised values. Thus, by not taking revisions into account, it is possible that we are substantially underestimating the marginal utility of including Twitter, so our results should be viewed as a lower bound on the true improvement. If we were to see the same effect of revisions in other countries, then we would expect the error to be decreased by a factor of 5.

We note that Twitter reduces mean squared error, but often worsens relative error. This means that Twitter reduces error more during weeks with large values than weeks with small values. It is not clear why this discrepancy exists, but this suggests that Twitter is most useful during weeks when it is most needed: during in-season and near-peak weeks where the rate is high. Our error metrics do not distinguish between in-season and off-season periods, but a finer-grained analysis is worth considering in future work, in light of these findings.

Distinguishing in-season and off-season weeks may also provide a more reflective estimate of a system's performance, because the performance is arguably less important off-season, depending on its application. More generally, mean error across all weeks is not necessarily the best metric for a surveillance system, although we use it here because it is a standard and interpretable metric. A recent CDC contest, for example, evaluated influenza forecasting systems by their ability to predict specific milestones, such as the peak and duration of a season (CDC 2013). Ultimately, the most appropriate metric for evaluation depends on the needs of practitioners.

Conclusion

We have presented, to the best of our knowledge, the most geographically comprehensive study to date evaluating the performance of Twitter-based influenza surveillance. We compared our Twitter estimates to ground truth data from ten countries, and found the Twitter data to be significantly correlated with all ten datasets. Additionally, we showed that Twitter data offers marginal utility over a standard autoregressive baseline, with increasing improvements with larger reporting lags. We analyzed the relationship between the system performance and characteristics of each location and found that Twitter surveillance tends to work better for more populous countries with higher Twitter activity.

Acknowledgements

Michael Paul is supported by a Microsoft Research PhD fellowship. Approved for public release LA-UR-14-27933.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.; and Liu, B. 2012. Twitter improves seasonal influenza prediction. In *International Conference on Health Informatics*.

Broniatowski, D. A.; Paul, M. J.; and Dredze, M. 2013. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE* 8(12):e83672.

Broniatowski, D.; Paul, M. J.; and Dredze, M. 2014. Twitter: Big data opportunities. *Science* 345(6193):148.

CDC. 2013. Announcement of requirements and registration for the Predict the Influenza Season Challenge. Technical Report FR 70303, 2013:70303-5, Centers for Disease Control and Prevention.

Location	Mean Squared Error (MSE)			Normalized MSE			Mean Relative Error		
	Base.	Tw.	Red.	Base.	Tw.	Red.	Base.	Tw.	Red.
	1-week lag ($\ell = 1$)								
Australia	0.92	0.83	10.5	0.080	0.072	10.5	0.116	0.116	-0.4
Canada	59.45	54.88	7.7	0.356	0.328	7.7	0.285	0.290	-1.7
England+Wales	7.67	7.71	-0.6	0.200	0.201	-0.6	0.274	0.290	-6.1
Ireland	34.47	33.92	1.6	0.132	0.130	1.6	0.347	0.419	-20.6*
New Zealand	244.99	200.98	18.0	0.215	0.176	18.0	0.362	0.449	-24.2
Northern Ireland	54.42	51.54	5.3	0.225	0.213	5.3	0.280	0.325	-15.8*
Scotland	20.68	21.35	-3.2	0.219	0.226	-3.2	0.262	0.307	-17.3*
South Africa	767.75	725.05	5.6	0.215	0.203	5.6	0.143	0.139	2.6
United States	0.13	0.11	15.3	0.129	0.109	15.3	0.050	0.050	0.3
Wales	13.34	12.97	2.7	0.608	0.591	2.7	0.358	0.393	-9.9
	2-week lag ($\ell = 2$)								
Australia	3.05	2.46	19.5*	0.265	0.214	19.5*	0.209	0.209	-0.1
Canada	77.33	58.72	24.1*	0.460	0.350	24.1*	0.280	0.295	-5.6
England+Wales	14.07	12.97	7.8	0.368	0.339	7.8	0.327	0.374	-14.4
Ireland	85.63	81.02	5.4	0.326	0.308	5.4	0.528	0.745	-41.3*
New Zealand	753.30	474.48	37.0	0.648	0.408	37.0	0.420	0.768	-82.8
Northern Ireland	101.82	95.57	6.1	0.420	0.394	6.1	0.347	0.423	-22.1*
Scotland	37.70	37.87	-0.5	0.398	0.400	-0.5	0.320	0.387	-20.9*
South Africa	1671.08	1383.37	17.2	0.488	0.404	17.2	0.197	0.186	5.3
United States	0.24	0.20	17.7	0.248	0.204	17.7	0.084	0.068	18.9*
Wales	15.49	14.43	6.8	0.708	0.660	6.8	0.390	0.433	-11.1
	3-week lag ($\ell = 3$)								
Australia	5.31	3.75	29.2*	0.462	0.327	29.2*	0.269	0.258	4.1
Canada	95.45	59.90	37.2*	0.564	0.354	37.2*	0.306	0.309	-1.0
England+Wales	24.75	22.41	9.5	0.645	0.584	9.5	0.406	0.517	-27.3*
Ireland	129.45	119.88	7.4	0.489	0.452	7.4	0.591	0.960	-62.5*
New Zealand	1210.43	494.07	59.2	1.031	0.421	59.2	0.540	0.636	-17.8
Northern Ireland	131.50	120.08	8.7	0.540	0.494	8.7	0.362	0.484	-33.7*
Scotland	50.17	48.05	4.2	0.529	0.507	4.2	0.335	0.451	-34.6*
South Africa	2842.66	2122.95	25.3	0.871	0.651	25.3	0.246	0.211	14.3
United States	0.36	0.24	33.6*	0.373	0.248	33.6*	0.109	0.083	23.5*
Wales	15.13	14.64	3.3	0.693	0.670	3.3	0.414	0.448	-8.3

Table 2: The raw evaluation metrics for the baseline nowcasting model (Base.), the model including Twitter data (Tw.), and the error reduction when using the Twitter model (Red.), which is the same value reported in Table 1. Because different locations report values on different scales, the mean squared error (MSE) is not comparable across locations. We therefore also report a normalized variant of MSE, which divides the MSE by the variance of the data: this is closely related to the standard r^2 metric, which is 1 minus the normalized MSE that we report. Finally, we show the mean relative absolute error: $|\hat{y}_w - y_w|/y_w$.

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

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

Dredze, M.; Paul, M.; Bergsma, S.; and Tran, H. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI HIAI Workshop*.

Dredze, M.; Cheng, R.; Paul, M.; and Broniatowski, D. 2014. Healthtweets.org: A platform for public health surveillance using twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*.

Eiji Aramaki, S. M., and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP*.

Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; and Smith, N. A. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Association for Computational Linguistics (ACL)*.

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.

Goel, S.; Hofman, J. M.; Lahaie, S.; Pennock, D. M.; and Watts, D. J. 2010. Predicting consumer behavior with web search. *PNAS* 107(41):17486–17490.

Han, B.; Cook, P.; and Baldwin, T. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49:451–500.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.

Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *NAACL*.

Lampos, V., and Cristianini, N. 2012. Nowcasting events from the social web with statistical learning. *ACM TIST* 3(4):72:1–72:22.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of Google flu: Traps in big data analysis. *Science* 343(6176):1203–1205.

Nagar, R.; Yuan, Q.; Freifeld, C. C.; Santillana, M.; Nojima, A.; Chunara, R.; and Brownstein, S. J. 2014. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res* 16(10):e236.

Nie, L.; Akbari, M.; Li, T.; and Chua, T.-S. 2014a. A joint local-global approach for medical terminology assignment. In *SIGIR 2014 Workshop on Medical Information Retrieval*, 24–27.

Nie, L.; Zhao, Y.-L.; Akbari, M.; Shen, J.; and Chua, T.-S. 2014b. Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE Transactions on Knowledge and Data Engineering* 99(PrePrints):1.

Oussalah, M.; Bhat, F.; Challis, K.; and Schnier, T. 2012. A software architecture for twitter collection, search and geolocation services. *Knowledge-Based Systems*.

Paul, M. J.; Dredze, M.; and Broniatowski, D. 2014a. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Paul, M. J.; Dredze, M.; and Broniatowski, D. A. 2014b. Challenges in influenza forecasting and opportunities for social media. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Santos, J. C., and Matos, S. 2014. Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling* 11(S1):S6.

Signorini, A.; Segre, A.; and Polgreen, P. 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.

Sun, X.; Ye, J.; and Ren, F. 2014. Real time early-stage influenza detection with emotion factors from Sina Microblog. In *5th Workshop on South and Southeast Asian NLP*.

Wing, B. P., and Baldridge, J. 2011. Simple supervised document geolocation with geodesic grids. In *ACL*.

Location	Source, URL, and Description
Australia	Department of Health `http://www.health.gov.au/internet/main/publishing.nsf/Content/cda-surveil-ozflu-flucurr.htm` ILI Sentinel taken from the Australian Influenza Surveillance Report figure 6: "Weekly rate of ILI reported from GP ILI surveillance systems" with the unit described as "Rate per 1,000 consultations". Weeks begin Monday.
Canada	Public Health Agency of Canada `http://www.phac-aspc.gc.ca/fluwatch/13-14/index-eng.php` ILI Sentinel taken from FluWatch Report figure 5: "Influenza-like-illness (ILI) consultation rates by report week" with the unit described as "Rate per 1,000 patient visits". Weeks begin Sunday.
Ireland	Health Protection Surveillance Centre `http://www.hpsc.ie/A-Z/Respiratory/Influenza/SeasonalInfluenza/Surveillance/InfluenzaSurveillanceReports/` ILI Sentinel taken from Influenza Surveillance Report figure 1: "ILI sentinel GP consultation rates per 10,000 population" with the unit described as "ILI rate per 100,000 population". Weeks begin Monday.
New Zealand	Institute of Environmental Science and Research `https://surv.esr.cri.nz/virology/influenza_weekly_update.php` ILI Sentinel taken from Influenza Weekly Update figure 2: "Weekly consultation rates for influenza-like illness in New Zealand, 2010-2014" with the unit described as "Consultation rate (per 100,000)". The Influenza Weekly Update only reports during the influenza season in New Zealand which typically lasts between weeks 18 to 44. Weeks begin Monday.
South Africa	National Institute of Communicable Diseases `http://www.nicd.ac.za/?page=surveillance_bulletin&id=15` Hospital consultation data taken from National Institute of Communicable Diseases Monthly Surveillance Bulletin. The unit measured is the number of private hospital outpatient consultations with a discharge diagnosis of pneumonia and influenza. Weeks begin Sunday.
United Kingdom	Public Health England `http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/SeasonalInfluenza/EpidemiologicalData/` ILI Sentinel taken from the National Influenza Report in the tables from the "Weekly consultation rates in national sentinel schemes" section. Weeks begin Monday.
United States	Centers for Disease Control and Prevention `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html` ILI Sentinal data from the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). The CDC coordinates the network and publishes weekly reports showing the percentage of outpatient consultations for ILI. National rates as well as rates for the 10 HHS regions are available. Weeks begin Sunday.

Table 3: Information about the surveillance data sources. For each country, we note the agency who provided the data, the URL from which the data were downloaded, and additional comments about the metrics and data availability. We also note which day of the week is the starting day for weekly counts.