# Validating Models for Disease Detection Using Twitter

Todd Bodnar
Pennsylvania State University
Department of Biology
University Park, PA 16802
tjb5215@psu.edu

Marcel Salathé
Pennsylvania State University
Department of Biology
University Park, PA 16802
salathe@psu.edu

## ABSTRACT

Data mining social media has become a valuable resource for infectious disease surveillance. However, there are considerable risks associated with incorrectly predicting an epidemic. The large amount of social media data combined with the small amount of ground truth data and the general dynamics of infectious diseases present unique challenges when evaluating model performance. In this paper, we look at several methods that have been used to assess influenza prevalence using Twitter. We then validate them with tests that are designed to avoid and illustrate issues with the standard k-fold cross validation method. We also find that small modifications to the way that data are partitioned can have major effects on a model's reported performance.

## Categories and Subject Descriptors

I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis

## General Terms

Algorithms, Reliability, Experimentation

## Keywords

data mining, regression, machine learning, Twitter

## 1. INTRODUCTION

The rapid adoption of social media and the internet in general has opened the door for novel developments in epidemiology [11, 7, 1, 12, 2, 13, 3, 5]. Much of this research has been aimed at data mining social media services such as Twitter or Facebook. Due to its openness, Twitter has been of particular interest [9, 15, 12, 4]. The site's microblogging and mobile communication features make it particularly useful for determining current levels of disease.

Given the rapid rise of social media usage, assessing disease prevalence using social media will become increasingly important. It is therefore prudent to continuously validate the underlying models [2, 1, 3]. Methods for validation assume that the training and testing data are independent of each other. While this assumption is never completely true, it is often sufficient. However, due to the strong spatial and temporal nature of infectious disease dynamics – along with a lack of multiyear social media datasets – this assumption may result in an inaccurate model.

In this paper, we take previously published models [3, 6, 13] and perform a battery of tests to check for potential issues. We do this by comparing the results of a traditional influenza related tweet dataset to a dataset of tweets that has not been filtered for a specific topic, a dataset of tweets related to a topic that is irrelevant to influenza, and a set of frequencies generated from random sine waves. In addition, we compare 10 fold and leave-one-out validation where the testing data are either from a different region or time than the training data. We find that (i) seemingly irrelevant tweets are moderately successful in assessing influenza prevalence, (ii) generated frequencies are often as good as measured frequencies from social media, and (iii) the choice of the validation method greatly affects the model's reported performance.

## 2. DATA SETS

### 2.1 Influenza Prevalence

The CDC defines ILI (influenza like illness) as an illness with a fever and a cough or sore throat without a known cause other than influenza. Because ILI is indistinguishable from influenza, except through expensive tests, most data is reported as ILI prevalence instead of influenza prevalence. We used the percentage of doctor's visits that were for ILI between October 2, 2011 and May 26, 2012, as reported by the CDC,[1] to serve as the ground truth. The CDC provides this data both on a national level and as a set of 10 HHS regions.

### 2.2 Tweets

We collected 238,506,796 tweets from the continental United States between October 2, 2011 and May 26, 2012 – a 34 week span – through Twitter's API. The tweets were acquired by requesting all tweets with high-resolution geospatial information within a bounding box that covers the continental United States. By limiting our requests to tweets with high-resolution geospatial data, we potentially introduced a bias in the data. However, this allowed us to avoid being rate limited by Twitter, guaranteeing that the dataset contained every tweet from Twitter, subject to the above parameters.

Each tweet consists of geospatial information, the time that the tweet was sent, and the contents of the message

[1] http://www.cdc.gov/flu/weekly/pastreports.htm

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
*WWW 2013 Companion*, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

| Y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| R | .87 | .88 | .63 | .91 | .98 | .95 | .95 | .98 | .89 | .90 |

**Table 1: Predicting region Y's ILI prevalence simply based on the other 9 regions' current prevalences with a multivariable regression illustrates the strong relationship between the regions' disease levels.**

tweeted, along with other information such as the user's profile picture and sign up date. The quality of the tweet's geospatial information varies greatly based on how the user sent it. For example, a tweet sent from a laptop may only have information from which city or state it originated from. In our case, we limited our search to tweets with longitude and latitude coordinates indicating that the tweets most likely came from gps equipped devices such as cell phones. A tweet contains at most 140 characters of text. Note that we did not limit our collection to tweets with a specific set of keywords.

We trained our models on 6 subsets generated from this data. The first subset was simply the entire dataset grouped by each week. The second subset was limited to tweets that contained at least one of the following ILI related keywords: 'flu', 'cough', 'fever', 'headache' or 'head ache'. We defined a third subset of the data using the keywords 'zombie', 'zed', 'undead' and 'living dead'. Since these keywords are presumably unrelated to ILI, this subset serves as a test for odd model behavior. The other three subsets are the same as the first three, but also divided based on which region the tweet came from. We used the 1000 most common words in each of the subsets as the list of keywords for the models. In the ILI dataset, this includes all of the words that were tweeted an average of at least one time per week. The other subsets were also of the top 1000 words to avoid biasing caused by a difference in the amount of data being fed into the models. We did not filter out stop words because, as mentioned by Culotta [3], stop words such as 'I' or 'have' provide valuable information if the tweets also have an ILI keyword. Because of daily fluctuations in Twitter use, all keyword trends are measured by their frequency.[2]

In addition to these 6 datasets, we 'simulate' keyword frequencies by generating another two datasets. We generate one-thousand sine curves with random wavelengths and add noise generated by a normal distribution with a standard deviation of 0.1 to each point. They are then divided by 1000 to be of the same scale as the actual frequencies. We add .001 to each point to avoid negative frequencies. We also generate one for regional data where the wavelengths are fixed across regions but the noise is not. As with the irrelevant tweet dataset, these serve as control groups.

# 3. MODELS

## 3.1 Regression on Tweet Count

Following previous work [6, 3], we first consider using a linear regression of the raw count of tweets that contain at least one of the keywords, as defined above, to predict the CDC's ILI prevalence:

$$logit(CDCRate) = \beta_0 + \beta_1 logit(x) + \epsilon \qquad (1)$$

[2]The datasets and associated code are available at http://github.com/salathegroup/w3cRio

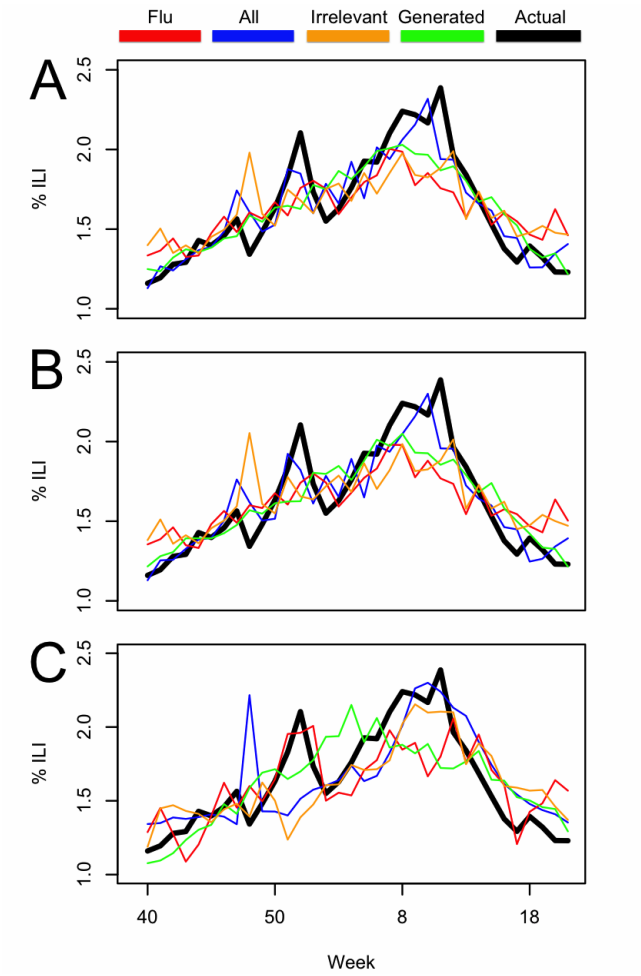


**Figure 1: Results from (a) SVM regression, (b) multivariable regression, and (c) single regression for each dataset compared to the CDC's national reported ILI levels during the 2011-2012 influenza season. Each data point is the result of a model trained on the other 33 week's data.**

Where $\beta_0$ and $\beta_1$ are coefficients, $\epsilon$ is the error function,$x$ is the number of Tweets containing at least one of the keywords and $logit(x) = log(x/(1-x))$.

## 3.2 Multivariable Regression

To gain more information from the tweets, we consider multivariable regression [3, 10].

$$logit(CDCRate) = \beta_0 + \sum_{i=1}^{n} \beta_i logit(x_i) + \epsilon \qquad (2)$$

Where $x_i$ is the frequency of the $i^{th}$ keyword.

## 3.3 Select Best Keyword

It has been argued that multivariable regression is prone to overfitting [6, 3]. An alternative solution to multivariable regression is to perform regression on the keyword that correlates the best with the training data, and use it for the regression model.

| Model | Flu | All | Irrelevant | Generated |
|---|---|---|---|---|
| Count | .4184 | .4344 | .0089 | .3529 |
| Multi | .7681 | .8774 | .6300 | .8367 |
| Best | .6946 | .6583 | .7991 | .7313 |
| SVMr | .7557 | .8580 | .7382 | .8766 |

**Table 2: Average correlation of the models' predictions and the CDC's national ILI prevalence.**

| Model | Flu | All | Irrelevant | Generated |
|---|---|---|---|---|
| Multi | .3493 | .6468 | .2860 | .2713 |
| Best | .1575 | .2381 | .3158 | .6653 |
| SVMr | .4538 | .7378 | .4270 | .7113 |

**Table 3: Mean correlation of the results of a model trained on 9 regions and evaluated on the last.**

### 3.4 SVM Regression

We consider a form of regression that utilizes a SVM (support vector machine) which has been shown to predict ILI prevalence well [13]. A SVM defines a multi-dimensional hyperplane that divides the training data. While this hyperplane is generally used in classification problems to divide two classes, it also allows for regression based on a sample's distance from the plane [14].

## 4. MODEL VALIDATION

Models are evaluated by dividing the dataset into training and validation sets. The way that the sets are divided has the potential to greatly affect the measured performance of the models [10]. Examples of these issues are

1. Too much data used in training results in few points to compare to the model's results.
2. Too little data used in training results in a poorly fitted model.
3. If the testing data is too similar to the training data, overfitting may not be detected.
4. If the testing data is too different from the training data, the model will perform badly regardless of it's sophistication.

As a concrete example of issue 3, consider the commonly used method of reserving one region's data for validation. As information about other regions gives a fair bit of information about a region (see table 1), there is a risk that a model will present good results in the testing data even if the model has not learned the system's underlying dynamics.

Aside from a simple percentage split, we allow for 10-fold-cross validation. In cross validation, the dataset is divided into $k$ equally sized splits. Each split is used to test a model that was trained on the remaining $k-1$ parts. In addition, we allow for leave-one-out cross validation where each datapoint is used to test a model that was trained by the remaining data.

## 5. RESULTS

We first evaluate the models with the national level data through leave-one-out validation (figure 1) and 10-fold-cross validation (table 2). In the case of 10 fold validation, we repeated the evaluation 100 times with different, randomly generated splits. With both validation methods, multivariable and SVM regression performed similarly. We corroborate Culotta's finding that multivariable regression performs better than regression on just the count of relevant tweets contrary to Ginsberg et al.'s findings in Google search queries. For a discussion on why this may be, see [3]. Because of its much lower performance, we ignore it for the rest of the analysis.

When we repeat this procedure on a regional level with each region being a 'fold', we observe similar behavior (table 3, fig 2). However the accuracies are lower in every case. This suggests that a model with what appears to be better performance may not necessarily be better than one with a lower level of performance if the first model's testing set was temporally separated while the second model's testing set was spatially separated.

It may appear that both multiple regression and SVM regression have similar accuracies in the regional data, however their results from the generated dataset are noticeably different. The intuitive conclusion would be that SVM regression performs better than multiple regression. This is not necessarily so. In the case of SVM regression, real Twitter data is barely a better predictor of ILI than generated sine curves. This calls into question the benefits of using social media with SVM regression, if randomly generated data performs nearly as well.

Interestingly, the dataset that was not filtered resulted in a higher correlation in 3 of the 4 models. This may be due to the filtering process removing potentially insightful tweets that do not contain any of the keywords. Another possibility is that reducing the number of tweets makes the data's trends more susceptible to random fluctuations and thus noisier.

## 6. CONCLUSIONS

In this paper we evaluated several well known regression models on their ability to accurately assess disease prevalence from tweets. We found that even irrelevant tweets and randomly generated datasets were able to assess disease levels comparatively well. This could serve as a ground level for evaluating other models: if a model can do only slightly better with seemingly relevant data than with seemingly irrelevant or random data, then it is probably not learning much from the tweets and its ability to fit the data can be attributed to other factors.

The ability for even randomly generated curves to fit the data may be explained by either spatial or temporal autocorrelation. For example in 5 fold cross validation, a model may simply interpolate between points in the dataset instead of gaining information from the tweets. Future work could look at other diseases that have less predictable long term dynamics, such as gastroenteritis or asthma. Another possibility is that – especially in the full dataset – tweets about other events that happened around the same time that ILI peaked could be chosen by the model as a predictor, but clearly this would not be expected to replicate across multiple years.

Finally, we found that the way that the training and testing data were divided had a strong effect on the reported performance of a model. Future work could build a mathematical model to explore these effects and develop a method

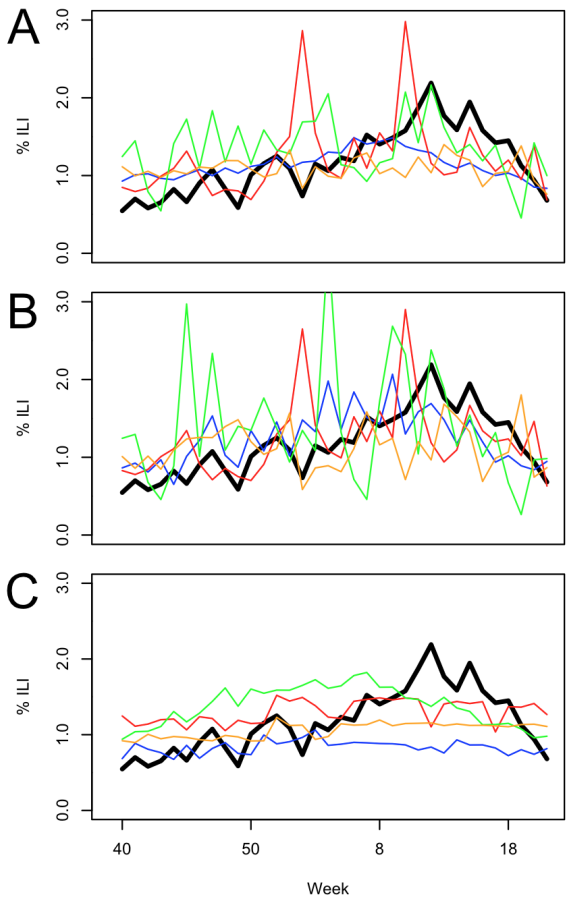


**Figure 2: As with figure 1, but results for region 10 from models trained on regions 1-9.**

to evaluate models in a way that best measures their true performance.

## 7. ACKNOWLEDGMENTS

Weka [8] and R [16] were used for most of the data analysis and visualization in this paper. MS acknowledges funding from a Branco Weiss fellowship.

## 8. REFERENCES

[1] D. Butler. When Google got flu wrong. *Nature*, 2013.

[2] H. A. Carneiro and E. Mylonakis. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564, Nov. 2009.

[3] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages . In *the First Workshop*, pages 115–122, New York, New York, USA, 2010. ACM Press.

[4] I. De la Torre-Díez, F. J. Díaz-Pernas, and M. Antón-Rodríguez. A content analysis of chronic diseases social groups on facebook and twitter. *Telemedicine journal and e-health: the official journal of the American Telemedicine Association*, 18(6):404–408, July 2012.

[5] A. F. Dugas, Y. H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman. Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases*, 54(4):463–469, Jan. 2012.

[6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, Feb. 2009.

[7] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17486–17490, Oct. 2010.

[8] M. Hall, E. Frank, G. Holmes, and B. Pfahringer. The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.*, 2009.

[9] C. Hawn. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health Affairs*, 2009.

[10] S. Marsland. Machine learning: an algorithmic perspective. 2009.

[11] M. Salathé, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani. Digital epidemiology. *PLoS computational biology*, 8(7):e1002616, July 2012.

[12] M. Salathé and S. Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199, Oct. 2011.

[13] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS one*, 6(5):e19467, 2011.

[14] A. Smola. Support vector regression machines. *Advances in neural information processing systems*, 1997.

[15] C. St Louis and G. Zorlu. Can Twitter predict disease outbreaks? *BMJ (Clinical research ed.)*, 344:e2353, 2012.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2012.