On the Ground Validation of Online Diagnosis with Twitter and Medical Records

Todd Bodnar*
Pennsylvania State University
Center for Infectious Disease Dynamics and Department of Biology
meme@psu.edu

Vicki Barclay

Conrad S Tucker
Pennsylvania State Univeristy
Department of Engineering Design and Industrial and Manufacturing Engineering
conrad.tucker@psu.edu

Marcel Salathé
Pennsylvania State University
Center for Infectious Disease Dynamics and Department of Biology
salathe@psu.edu

ABSTRACT
This is an abstract

Categories and Subject Descriptors
I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and Science*

General Terms
Experimentation, Validation

Keywords
Twitter, Validation, Digital Epidemiology, Remote Diagnosis

1. INTRODUCTION
Disease surveillance systems – traditionally relying on reports from medical practitioners – are an important part of disease control. However, these traditional surveillance systems are often costly and slow to respond[8, 3, 13]. The widespread adoption of the Internet by the general public has provided opertunities for the development of novel disease surveillance methods. Compared to traditional systems, where data is provided by medical diagnosis, these new systems provide either semi-automatic – through long term self reporting systems[10, 15] – or fully automatic – through datamining search queries or social media – disease surveillance. While these methods are cheaper, faster and cover a larger number of individuals than traditional systems, one can be less confident about their results than the results from a system based on professional diagnosis. In this paper, we develop a system that performs long term surveillance on Twitter users with methods trained on professionally diagnosed data that combines the advantages of all three of these previous systems.

Previous work with datamining social media has focused on methods to replicate the patterns found in traditional surveilance networks[1, 4, 6]. However, these methods have several limitations. First, they generally do not differentiate between an individual with an illness and an individual that is worried about an illness; which may have resulted in a predicted influenza rate that was much higher then the actual 2013 influenza rate [1, 2, 11, 9]. Second, these methods cannot be extended to areas without a previous surveillance network. Finally, these methods are fundamentially incapapble of detecting diseases that do not show strong spatio-temporal patters such as mental illness, obesity or Parkinson's disease[]. Instead of top-down methods to measure levels of disease in a population, we approach this problem from the bottom-up. This addresses all three of these issues: we only diagnose individuals that are likely to have the disease, and not just interested in the disease; we do not require previous data when applying these methods to new problems or locations; and these methods can easily generalize to diseases that do not show strong patters because we focus on an individual by individual level.

Systems, such as Influenzanet or Flu Near You, use self-reported symptoms to diagnose an individual also work from a bottom-up approach.[10, 15] These systems have the potential to be better than traditional surveillance systems because they update in near-real-time and can detect cases even when the user has not gone to their doctor. These systems require the user to sign up which allows for long term studies which are not normally able to be done with Tweets or search queries. However this reduces the number of users studied compared to datamining approaches. For

*Corresponding author

example, Flu Near You had a total of 9,456 users report during the week ending 29 December 2013. Marquet et al. [10] has shown a large drop out rate with only 53% users participating for five or more weeks. While this amount of data is sufficient for many purposes, a system which is based on Twitter's millions of active users would open the door to novel questions.

We develop this such a system as follows. In section 2 we describe the collection of professional diagnosis of Twitter users and collect their Twitter information. In section 3 we consider extracting textual information from Tweets as a method for diagnosing influenza. In section 4 we consider anomalies in a user's Tweeting behavior as a signal for diagnosing influenza. In section 5 we extend these methods to other users on a persons social network to diagnose the original person. In section 6 we aggregate the results of the previous classifiers to develop a more accurate meta-classifier.

2. DATA COLLECTION

2.1 Medical records

We recieved information about 104 individuals from the Pennsylvania State University's Health Services that were professionally diagnosed with Influenza by a medical professional during the 2012-2013 Influenza season (see figure 1). Additionally we collect information from 125 individuals that were *not* diagnosed as a control set. The participants were mostly students (72% were between the ages of 18 and 22) and slightly female ($133/226 \approx 58.8\%$.) Data collection was approved through the Pennsylvania State University's IRB (#41345.) Additionally we recieved the Twitter handles from 119 of these individuals.

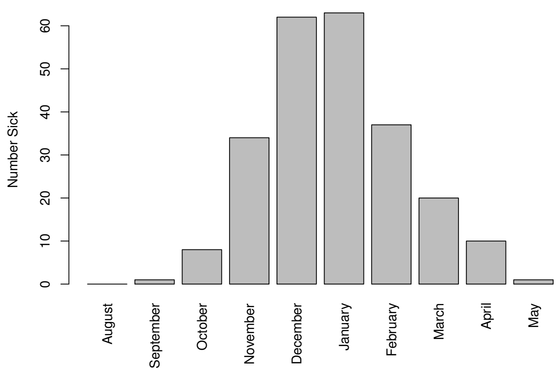

Figure 1: The rate of professionally diagnosed Influenza cases during the 2012-2013 season in our sample.

2.2 Twitter records

While we received a total of 119 Twitter accounts, 15 were discarded because the associated accounts were either nonexistent, banned, or private. For each of the remaining 104 accounts, we pulled their profile information, their friends and followers information, their most recent 3000 tweets, and their friends profiles and tweets. Some users did not tweet during the month that they were sick; we kept those accounts as part of the control group. We were limited to the most recent 3000 tweets by Twitter's time line query, but this only effected two accounts – both of which posted multiple times per hour and were thrown out because we could only look back a few days.

We collected data by calling the Twitter API on the user account that we queried the longest time ago. Tweets, profile and follower information queries have separate rate limits and were collected in parallel. The 104 seed accounts collected above were given higher priority over their friends and followers. In total, we collected 37,599 tweets from the seed accounts and 30,950,958 tweets from 913,082 accounts that they either followed or were followed by.

3. TEXT BASED SIGNALS

In this section, we consider diagnosis based on classifying an individual tweet's content as either about ILI or not. We begin by dividing the tweets into two sets: tweets that were posted the same month that a user was sick, and tweets that were posted other times. We find a total of 1609 tweets from 35 users in the first category.

First we go the route of defining a set of keywords that are positive signals of influenza. We chose {flu, influenza, sick, cough, cold, medicine, fever} as our set of keywords. Of these seven keywords, we find a significant effect in 6 of the keywords during months when the user had ILI. (See table 1). Additionally, we try algorithmically selecting keywords by first finding the 12,393 most common keywords in the data set. We then rank them based off of information gain and choose the top 10, 100 or 1000 keywords from the list. In both of these cases, we preprocess the data by tokenizing the text on the characters ".,;':"()?!" - as well as spaces, tabs and line breaks - remove stop words[1], perform Porter stemming [12, 16] and convert the text to lower case. We then use the occurance or absence of these keywords as features for classification. We use naive bayes, random forest, J48, logistic regression and support vector machines to classify a user as being sick in a given month or not (see figure 2).

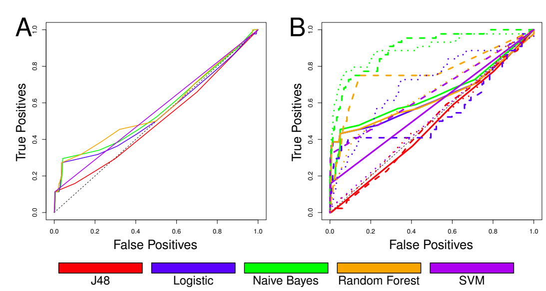

Figure 2: The ROC of classifiers that use hand chosen keywords (a) and algorithmically chosen keywords (b) to determine if an individual is ill. The top 10 (solid line), 100 (dashed line) and 1000 (dotted line) were selected as the features.

Additionally, we hand rate all 1609 tweets, that were posted by individuals during the time of their illness, for information regarding the user's health. We also sample a randomly selected set of 1609 tweets from times when the users did not have ILI as a control. We find 58 tweets from 17

[1] Stop words taken from Weka's stoplist version 3.7.10.

Word	Total	Odds Ratio	Significance
flu	25	40.14	<0.0001
influenza	1	0.00	0.8325
sick	128	5.22	<0.0001
cough	18	4.48	0.0094
cold	82	1.45	0.4154
medicin	9	11.20	<0.0001
fever	13	26.20	<0.0001

Table 1: Keyword effects.

Sick	Not Sick	
17	18	Sick
0	66	Not Sick

Table 2: Confusion matrix of a Tweet-Classification based diagnosis system. Rows are of true values, columns are of predicted values.

(17/35 = 48.57%) individuals in our study that are about the user being sick. We also find zero tweets about ILI during times when they did *not* have ILI. While the use of a "human" classifier clearly does not scale, it allows for an approxamatly 100.0% accurate classification. Since regular machine learning algorithms preform much worse than 100.0% accuracy, the human classifier gives us an upper limit to the accuracy of a health monitoring system based off of tweet classification . (See table 2)

4. FREQUENCY BASED SIGNALS

It is likely that a user tweets at a different rate when she is ill than she normally does. To detect this, we perform one-dimensional anomaly detection on each user's monthly tweeting rate as follows. First, we calculate the number of tweets in each month in the study period and discard any months where the user tweets less than ten times. This avoids issues caused by the user starting or stopping to use Twitter. We then calculate the z-score of the tweeting rate of the month that the user is ill by

$$z = \frac{|x - \bar{x}|}{\hat{s}} \tag{1}$$

Where $\bar{x}$ and $\hat{s}$ are the estimated mean and standard deviation of the user's tweeting rate. [7] We repeat this process for months when the user is not sick. We then decide that the user is sick if $z > 1.411$ where 1.411 was chosen through leave one out cross validation. We find a significant difference between the z-scores for months when a user was had ILI and months when the user did not ($p = 0.01303$, two-sample Kolmogorov-Smirnov test). Most of the time individuals are not sick (219 / 258 = 84.88%), resulting in a highly biased sample. Thus we optimize based on the F_1 score instead of accuracy. The optimal z-score cutoff results in $F_1 = 35.0\%$. (See table 3.)

5. NETWORK BASED SIGNALS

Even if a user is not currently active at all on Twitter, her friends or followers may give clues to her health status. We consider all text that a user's friends or followers tweeted

Sick	Not Sick	
14	25	Sick
27	192	Not Sick

Table 3: Confusion matrix of the classifier based on anomalous tweeting rates. Rows are of true values, columns are of predicted values.

and perform simular keyword analysis as was applied to the user's tweets in section 3. Because the number and activity of friends or followers greatly varies regardless of a user's health status, we normalize the counts here by the total number of characters her friends or followers tweeted. We find that most of the tested classifiers are able to detect a signal in both the user's followers' and friends' streams. (See figure 3.)

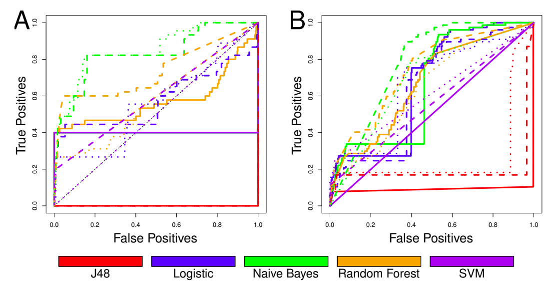

Figure 3: The ROC of classifiers based off Tweets from (a) accounts that follow a user and (b) accounts that a user follows. Line coloring and style are equivalent to figure 2.

6. META CLASSIFIER

So far we have considered five seperate methods for detecting illness based off of a user's Twitter activity. However, there is no reason that we cannot combine these methods to get a stronger signal. For example, while mining the user's text is the best of the five methods, she may stop tweeting while sick, which would be detected by the frequency-based anomaly classifier. Aggregating multiple classifiers by a 'meta-classifier' has been shown to be an effective method for increasing classification accuracy. [5, 14]

We start by selecting the classifier from each of the previous five approaches that has the largest area under the ROC curve (see figure 4 A). We then use the predicted distributions from these classifiers as the feature vector for the meta classifier. We use AdaBoost, bayesian classification, J48 decision trees, logit boost, and weighted voting to evaluate the meta-dataset. We then evaluate these methods with leave-one-out cross validation and see an increase in ROC area (and accuracy) over the best individual classifier (see figure 4 B).

7. CONCLUSIONS

In this paper, we have shown that it is possible to diagnose an individual based off of her social media data with high accuracy. Computational approaches to aid in disease diagnosis has been approached before, however they have been developed with a medical setting in mind. That is, the problem addressed was "can we diagnose an individual

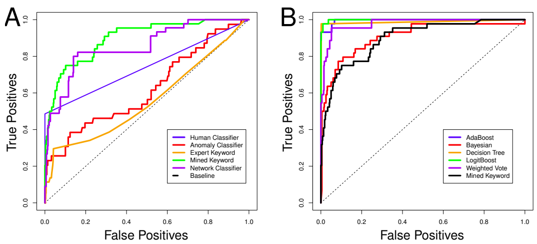

Figure 4: The accuracy of the previous classifiers (a) and the accuracy of various classifiers that use the previous classifier's results as features (b).

based off data gathered from medical tests run on her?" instead of "can we diagnose an individual solely based off of publically available social media data?" While we focus on the relativly benign case of remotely reconstructing a confidental diagnosis of Influenza, these methods could also be applied to stigmatized diseases, such as HIV, where being able to determine if an individual is HIV positive without her knowledge and with only her Twitter handle could result in serious social or economic effects. Half of the users explicitly stated that they were sick, and we were able to confidently determine illness in the other half of the cases through their data. It would seem that avoiding discussing a illness is not enough to hide one's health in the age of big data.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] T. Bodnar and M. Salathé. Validating Models for Disease Detection Using Twitter. *WWW 2013 Companion*, June 2013.

[2] D. Butler. When Google got flu wrong. *Nature*, 494(7436):155–156, Feb. 2013.

[3] E. H. Chan, T. F. Brewer, L. C. Madoff, M. P. Pollack, A. L. Sonricker, M. Keller, C. C. Freifeld, M. Blench, A. Mawudeku, and J. S. Brownstein. Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences*, 107(50):21701–21706, 2010.

[4] 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.

[5] D. Frossyniotis, K. S. Nikita, and A. Stafylopatis. A meta-classifier approach for medical diagnosis. *... and Applications of ...*, 2004.

[6] 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.

[7] F. E. Grubbs. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21, 1969.

[8] D. L. Heymann and G. R. Rodier. Hot spots in a wired world: {WHO} surveillance of emerging and re-emerging infectious diseases. *The Lancet Infectious Diseases*, 1(5):345 – 353, 2001.

[9] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. *cs.jhu.edu*, 2013.

[10] R. L. Marquet, A. Bartelds, S. P. van Noort, and e. al. Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003–2004 influenza season. *BMC Public ...*, 2005.

[11] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS computational biology*, 9(10):e1003256, Oct. 2013.

[12] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[13] 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.

[14] L. Todorovski and S. Džeroski. Combining classifiers with meta decision trees. *Machine Learning*, 50(3):223–249, 2003.

[15] S. P. Van Noort, M. Muehlen, and A. H. Rebelo. Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. ... = *European ...*, 2007.

[16] P. Willett. The Porter stemming algorithm: then and now. *Program: electronic library and information systems*, 2006.

APPENDIX

Keyword	Ratio	cont.	
flu	34.424	walk	3.077
health	11.360	childr	6.820
sick	5.019	incred	6.820
track	10.952	meal	6.820
stud	3.508	longer	6.820
asshol	9.090	succes	26.765
ton	9.090	accis	26.765
particip	20.667	holida	26.765
salt	20.667	luv	26.765
recov	40.118	oblig	26.765
fuck	2.963	path	26.764
sham	13.64	pract	26.764
row	10.180	prayer	26.765
win	2.947	reserv	26.765
rt	3.077	riot	26.765

Table 4: The thirty keyword stems with the highest positive predictive power ranked by significance. Ratio is calculated as the rate of occurance when a user is sick over the rate when a user is not sick.

A. KEYWORD RECOMENDATIONS

While our system should be trusted more than one based simply off of aggregated tweets, it is more computationally intensive than simply pulling data from a keyword stream.

These systems require the user to select a specific set of keywords before data collection can begin. Keywords representing symptoms such as “flu”, “cough”, “sore throat”, and “headache” are often chosen. We suggest that the thirty[2] keywords with the highest postivie predictive value (see table 4) be chosen as the parameters for a keyword stream. In addition to keywords related to symptoms (e.g. “flu” or “sick”) we also find keywords related to treatments (e.g. “health,” “prayer” or “recovery”) and keywords related to negative mood (e.g. vulgarities) to be more commonly tweeted when a user is ill.

[2]The Twitter API limits queries to thirty keywords.