

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
tjb5215@psu.edu
Conrad S Tucker

Vicki Barclay

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

Digital epidemiology, datamining Internet records to approach epidemiological questions in novel ways, has recently SOMETHING.

Previous work has generally focused on methods to replicate the patterns found in traditional surveillance networks. However, these methods have several limitations. First, they cannot generally differentiate between an individual with an illness and an individual that is worried about an illness. This may have resulted in a predicted influenza rate that was much higher than the actual 2013 influenza rate. Second, these methods cannot be extended to areas without a previous surveillance network. Finally, these methods are

^{*}Corresponding author

fundamentally incapable of detecting diseases that do not show strong spatio-temporal patterns such as mental illness, obesity or Parkinson's disease.

We present a system for online disease diagnosis through Twitter that has been trained on individuals that had been previously diagnosed by medical professionals. 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 patterns because we focus on an individual by individual level.

We develop this diagnosis system as follows. In section 3 we describe the collection of diagnosis and Twitter information. In section 4 we consider extracting textual information from Tweets as a method for diagnosing influenza. In section 5 we consider anomalies in a user's Tweeting behavior as a signal for diagnosing influenza. In section 6 we extend these methods to other users on a person's social network to diagnose that person. In section 7 we aggregate the results of the previous classifiers to develop a more accurate classifier.

2. RELATED WORK

People with issues [1, 2, 6] also plos paper!

Most work to this point considers finding messages in tweets (i.e. "I'm sick") or in keyword frequencies.

Keyword [3, 4]

Tweet classification [3, 6, 8]

3. DATA COLLECTION

3.1 Medical Records

3.2 Twitter Records

We received a total of 119 user accounts for Twitter from the survey; 15 of which were discarded because the associated accounts were either non-existent, 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 XX tweets from 913,082 accounts that they either followed or were followed by.

4. 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 words¹, perform Porter stemming [7, 9] and convert the text to lower case. We then use the occurrence 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 1).

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. Additionally we 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 (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

¹Stop words taken from Weka’s stoplist version 3.7.10.

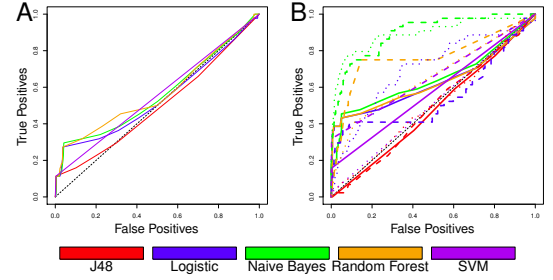


Figure 1: The ROC of classifiers that use hand chosen keywords (a) and algorithmically chosen keywords (b) to determine if an individual is ill. The top 10 (solid line), 100 (dashed line) and 1000 (dotted line) were selected as the features.

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.

tweet classification . (See table 2)

5. FREQUENCY BASED SIGNALS

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

$$z = \frac{|x - \bar{x}|}{\hat{s}} \quad (1)$$

Where \bar{x} and \hat{s} are the estimated mean and standard deviation of the user’s tweeting rate. [5] 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

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.

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.

a highly biased sample. Thus we optimize based on the F_1 score. The optimal z-score cutoff results in $F_1 = 35.0\%$. (See table 3.)

6. NETWORK BASED SIGNALS

Preliminary idea. Cascade effects causing echoes on social network. Also consider friends becoming ill around same time. Check @ tag

7. META CLASSIFIER

Combine features based off of previous signals, get something that's – hopefully – more accurate.

To this point, we have considered five approaches to detecting illness – keyword detection, keyword mining, human classification, frequency-based anomaly detection and network based classifiers – independently. Now we consider combining the results of the previous classifiers and develop a meta-classifier. (See figure 2) Concretely, we begin by selecting the classifier from each approach that has the largest area under the ROC curve. Then we use....

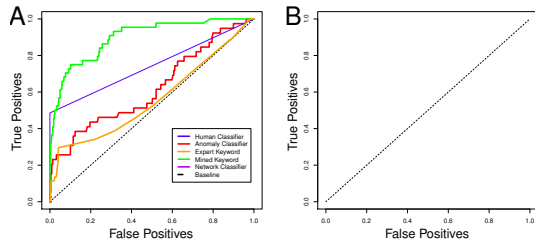


Figure 2: The accuracy of the previous classifiers (a) and the accuracy of various classifiers that use the previous classifier's results as features (b).

8. CONCLUSIONS

Comment on ~ 250 mill active T users \rightarrow 75 mill active US users, possibility of detecting $\sim 10\%$ anual ILI rate, \sim half are detectable on twitter Twitter, so 12.5 mill cases as upper estimate.

In this paper, we have shown that it is possible to diagnose an individual based off of her social media data with high accuracy. Computation 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 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 relatively benign case of remotely reconstructing a confidential 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.

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] 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] 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.
- [5] F. E. Grubbs. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21, 1969.
- [6] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. *cs.jhu.edu*, 2013.
- [7] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.
- [8] 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.
- [9] P. Willett. The Porter stemming algorithm: then and now. *Program: electronic library and information systems*, 2006.