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

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

Maybe Conrad, Maybe Vicky

Marcel Salathé
Pennsylvania State University
Department of Biology
University Park, PA 16802
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 → novel disease detection mechanisms.

Validation of this idea is important, but not done.

Pull med info of individuals professionally diagnosed with ILI and their twitter accts. Compare old methods. Suggest some new things.

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, 7]

3. DATA COLLECTION
3.1 Medical Records

*Corresponding author

3.2 Twitter Records
We received a total of 119 user accounts for Twitter from the survey; 15 of which were discarded because the screen names 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 YY accounts that they either followed or were followed by.

4. SIGNAL DETECTION
4.1 Event 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 AUTHOR and AUTHOR by 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 significantly different amount of 999 of the keywords during months when the user had ILI. (See table X). Additionally, we use METHOD to automatically find keywords with a significant effect (See table Y). In both of these cases, we preprocess the data by tokenizing the text with the regex "REGEX", remove stop words[1], perform iterated levins stemming and ignoring case. For each keyword x, we define an individual as sick on month m if their Twitter stream contains x at least one time. We find this method to correctly classify users % of the time. (See tables 2 and 3)

[1]Stop words taken from HERE

Word	Total	Odds Ratio	Significance
flu	25	40.13898	2.707071e-49
influenza	1	0	0.8325418
sick	128	5.224356	5.366579e-16
cough	18	4.479715	0.009395649
cold	82	1.453834	0.4154262
medicin	9	11.19795	1.654012e-05
fever	13	26.19746	1.030862e-18

Table 1: Keyword effects.

Sick	Not Sick	
1	2	Sick
3	4	Not Sick

Table 2: Confusion matrix of a classifier based on keywords from a domain expert. Rows are of true values, columns are of predicted values.

Finally, 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. Thus we would expect a health monitoring system based off of tweet classification to operate with at most 82.18% accuracy. (See table 4)

4.2 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

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

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	
1	2	Sick
3	4	Not Sick

Table 3: Confusion matrix of a classifier based on keywords derived from an algorithmic aproach. Rows are of true values, columns are of predicted values.

Sick	Not Sick	
17	18	Sick
0	66	Not Sick

Table 4: 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 5: 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 5.)

4.3 Network Based Signals

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

5. META CLASSIFIER

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

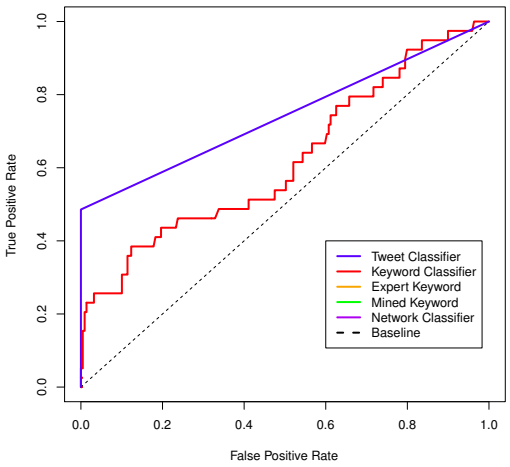

Figure 1: The accuracy of the previous classifiers (a) and the accuracy of various classifiers that use the previous classifier's results as features (b).

6. ANALYSIS

7. CONCLUSIONS

Comment on $\sim$ 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.

8. 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. 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.