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, 5] 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, 5, 6]

3. DATA COLLECTION

3.1 Medical Records

3.2 Twitter Records

Screen names taken from med records, total of 119 individuals gave us twitter handle. Threw out 15 accounts because either not legit, banned, or blocked for total of 104 seed accounts.

Tweets collected through twitter timeline query. API limits to most recent 3000 tweets per account. Two cases where this was an issue, both thrown out (users that consistently tweet multiple times per hour, barely any back data available. oddly twitter site does not give accurate tweet count.) Simple loop through every account. Once first pass, process by pulling any new tweets from user that has longest time since last query.

Pulled all friends / followers from accounts. Repeated tweet pull from these. Again, process oldest first.

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 (out of Y) tweets from X 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 choose KEYWORD LIST as our set of keywords. Of the 999 keywords, we find a significantly higher amount of 999 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 atleast one time. We find this method to correctly classify users % of the time. (See tables 1 and 2)

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

[*] Corresponding author

[1] Stop words taken from HERE

Sick	Not Sick	
1	2	Sick
3	4	Not Sick

Table 1: Confusion matrix of a classifier based on keywords from a domain expert. Rows are of true values, columns are of predicted values.

Sick	Not Sick	
1	2	Sick
3	4	Not Sick

Table 2: Confusion matrix of a classifier based on keywords derived from an algorithmic aproach. Rows are of true values, columns are of predicted values.

did not have ILI as a control. We find 58 tweets from 17 (17/119 = 14.29%) of the individuals in our study that are about them being sick. We also find x tweets from y/Z of individuals 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 at % accuracy. (See table 3)

4.2 Frequency Based Signals

Look at changes in behaviour based on illness. May have signal even if no relevant messaging.

In each user, take months when they did tweet, apply normalization (val - min)/(max-min)

Build distributions of months before, during, and after illness. Compare distributions. Try paired / unpaired

Fail to find sig difference between three sets → comment on benefits of this.

Try anomoly detection...

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. ANALYSIS

6. CONCLUSIONS

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

[6] 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.

Sick	Not Sick	
17	102	Sick
3	4	Not Sick

Table 3: Confusion matrix of a Tweet-Classification based diagnosis system. Rows are of true values, columns are of predicted values.