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!

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.

*Corresponding author

4. SIGNAL DETECTION
4.1 Event Based Signals
Most work to this point considers finding messages in tweets (i.e. "I'm sick") or in keyword frequencies.

Consider keywords, we try using an unsupervised method to find words that seperate months with and without ILI.

Now we try using hand chosen keywords, see if we can distinguish between individuals' sick and healthy months.

Finally, we hand rate all tweets in sick months to see if we can find tweets announcing illness. Find % of individuals do tweet about being sick. Also test subsample of non-sick months to see for false positives.

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.