Data Mining Social Media for Public Health Applications

Henry Kautz
University of Rochester
Rochester, NY 14610

Keywords: data mining, social media, healthcare

The online population creates a vast organic sensor network composed of individuals reporting on their activities, their social interactions, and the events around them. This firehose of data streams in real time, and is often annotated with context including GPS location, relationships, and images.

There is much activity in data mining social media for marketing campaigns (Richardson and Domingos 2002; Chen *et al.* 2010; Kirkpatrick 2012), financial prediction (Asur and Huberman 2010; Bollen and Mao 2011), and similar purposes. Recently, however, a smaller group of researchers have begun to leverage this sensor network for a singular public good: *modeling public health at a population scale*. Researchers have shown, for example, that Twitter postings can be used to track and predict influenza (Krieck *et al.* 2011; Signorini *et al.* 2011; Sadilek *et al.* 2012a; 2012b; Sadilek and Kautz 2013) and detect affective disorders such as depression (Zhang *et al.* 2010; Choudhury *et al.* 2013a; 2013b). Such work provides strong evidence that there is a strong health "signal" in social media.

Krieck *et al.* (2011) explored augmenting the traditional notification channels about a disease outbreak with data extracted from Twitter. By manually examining a large number of tweets, they showed that self-reported symptoms are the most reliable signal in detecting if a tweet is relevant to an outbreak or not. Researchers have also tried capturing the overall *trend* of a particular disease outbreak, typically influenza, by monitoring social media (Culotta 2010; Lampos *et al.* 2010; Chunara *et al.* 2012). Other researchers focus on more detailed modeling of the *language* of tweets and their relevance to public health in general (Paul and Dredze 2011) and to influenza surveillance in particular (Collier *et al.* 2011). Paul and Dredze developed a variant of topic models that captures the symptoms and possible treatments for ailments, such traumatic injuries and allergies, that people discuss on Twitter.

In our own project, *FluTracker*, we first automatically detected Twitter messages that suggest the author has the flu Sadilek *et al.* (2012a). We then constructed a probabilistic model that can predict if and when an individual will fall ill with high precision and recall on the basis of his social ties and co-locations with other people, as revealed by their Twitter posts (Sadilek *et al.* 2012b). Finally, we quantified the impact of social status, exposure to pollution, interpersonal interactions, travel patterns, and other important lifestyle factors on health using a unified statistical model (Sadilek and Kautz 2013).

Techniques similar to those developed for modeling infectious disease can be applied to study mental health disorders, such as depression, that have strong contagion patterns as well. Twitter has been used to monitor the seasonal variation in affect around the globe (Golder and Macy 2011). Choudhury *et al.* (2013a) examined patterns of activity, emotional, and linguistic correlates for childbirth and the postnatal course, and showed that mothers at risk for postpartum depression can be distinguished by linguistic changes captured by shifts in a relatively small number of words in their social media posts.

The work briefly described above are just first steps in data mining social media social media for public health, and they by and large make use of models and algorithms that are well developed in the AI community. The size and importance of public health applications, however, will also drive fundamental research on scalable machine learning and knowledge representation. Example tasks that require new algorithms and representations include:

- Learning dynamic relational models of health states, which generalize classical epidemiological models but support individual as well as aggregate predictions.

- Generalizing language, behavior, and network models across a range of physical and emotional health conditions.

- Developing methods for causal analysis in order to discover and measure global-scale influences on health.

This new approach to collecting and analyzing health information has the potential to revolutionize public health, by making detailed data about health, behavior, social structure, and geographic influences available in real time and at almost no cost.

References

S. Asur and B.A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 492–499, 2010.

Johan Bollen and Huina Mao. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10):91–94, 2011.

Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pages 1029–1038, 2010.

M. De Choudhury, S. Counts, and E. Horvitz. Major life events and behavioral markers in social media: Case of childbirth. In *16th ACM Conference on Computer Supported Cooperative Work (CSCW 2013)*, 2013.

M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*, 2013.

R. Chunara, J.R. Andrews, and J.S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.

N. Collier, N.T. Son, and N.M. Nguyen. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2(Suppl 5):S9, 2011.

A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, pages 115–122. ACM, 2010.

S.A. Golder and M.W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

David Kirkpatrick. Social media marketing: Data mining twitter for trends, sentiment and influencers, 2012. http://sherpablog.marketingsherpa.com/, retrieved 15 Dec 2012.

M. Krieck, J. Dreesman, L. Otrusina, and K. Denecke. A new age of public health: Identifying disease outbreaks by analyzing tweets. *Proceedings of Health WebScience Workshop, ACM Web Science Conference*, 2011.

V. Lampos, T. De Bie, and N. Cristianini. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases*, pages 599–602, 2010.

M.J. Paul and M. Dredze. A model for mining public health topics from Twitter. Technical report, Johns Hopkins University, 2011.

Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*, pages 61–70, 2002.

Adam Sadilek and Henry Kautz. Modeling the impact of lifestyle on health at scale. In *Sixth ACM International Conference on Web Search and Data Mining*, 2013.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

A. Signorini, A.M. Segre, and P.M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5), 2011.

Yuan Zhang, Jie Tang, Jimeng Sun, Yiran Chen, and Jinghai Rao. MoodCast: emotion prediction via dynamic continuous factor graph model. In *IEEE 10th International Conference on Data Mining (ICDM 2010)*, pages 1193–1198, 2010.