



JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Department of Biostatistics

BIOSTATISTICS SEMINAR

Bayesian Models for Mining Public Health Information from Twitter

Mark Dredze

Assistant Research Professor, Computer Science at Johns Hopkins University and
Research Scientist in the Human Language Technology Center of Excellence (HLTCOE).

Abstract

Twitter and other social media sites contain a wealth of information about populations and has been used to track sentiment towards products, measure political attitudes, and study social linguistics.

In this talk, we investigate the potential for Twitter to impact public health research. Specifically, we consider population surveillance, a major focus of public health that typically depends on clinical encounters with health professionals to collect patient data.

Individual users often broadcast salient health information, such as "sick with this flu fever taking over my body ughhhh time for tylenol", which indicates that not only does this person have the flu, but also a fever and is self-medicating with tylenol. Aggregating such content across millions of users could provide information about numerous aspects of illnesses in the population.

In this work we present the Ailment Topic Aspect Model (ATAM), a new Bayesian graphical model for Twitter that associates symptoms, treatments and general words with diseases (ailments.) When applied to 1.6 million health related tweets, ATAM discovers descriptions of diseases in terms of collections of words (symptoms and treatments) and partitions messages based on the referenced disease. The model discovers diseases corresponding to influenza, infections, obesity, insomnia, and several others. Furthermore, we demonstrate the effectiveness of this model at several tasks: tracking illnesses over times (syndromic surveillance), measuring behavioral risk factors, localizing illnesses by geographic region, and analyzing symptoms and medication usage. We show quantitative correlations with public health data and qualitative evaluations of model output. Our results suggest that Twitter has broad applicability for public health research.

**The Johns Hopkins Bloomberg School of Public Health
Department of Biostatistics, Wednesday, December 7, 2011
Room W2030 School of Public Health, 4:00-5:00pm (Refreshments: 3:30)**

For disability access information or listening devices, please contact the [Office of Support Services](#) at 410-955-1197 or on the Web at www.jhsph.edu/SupportServices. EO/AA



Bio:

Mark Dredze is an Assistant Research Professor in Computer Science at Johns Hopkins University and a research scientist in the Human Language Technology Center of Excellence (HLTCOE). He is also affiliated with the Center for Speech and Language Processing (CLSP) and is part of the Machine Learning Group. His research in natural language processing and machine learning has focused on graphical models, semi-supervised learning, information extraction, large-scale learning, speech processing and health informatics. He obtained his PhD from the University of Pennsylvania in 2009.

