ISDS 2012 Conference Abstracts

Detecting the Determinants of Health in Social Media

Caitlin Rivers*[1], Bryan Lewis[1] and Sean Young[2]

[1]Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Blacksburg, VA, USA; [2]UCLA David Geffen School of Medicine, Berkley, CA, USA

Objective

Create an analysis pipeline that can detect the behavioral determinants of disease in the population using social media data.

Introduction

The explosive use of social media sites presents a unique opportunity for developing alternative methods for understanding the health of the public. The near ubiquity of smartphones has further increased the volume and resolution of data that is shared through these sites. The emerging field of digital epidemiology[1] has focused on methods to analyze and use this "digital exhaust" to augment traditional epidemiologic methods. When applied to the task of disease detection they often detect outbreaks 1-2 weeks earlier than their traditional counterpart [1]. Many of these approaches successfully employ data mining techniques to detect symptoms associated with influenza-like illness [2]. Others can identify the appearance of novel symptom patterns, allowing the ability to detect the emergence of a new illness in a population [3]. However, behaviors that lead to increased risk for disease have not yet received this treatment.

Methods

We have created a methodology that can detect the behavioral determinants of disease in the population. Initially we have focused on risky behaviors that can contribute to HIV transmission in a population, however, the methodology is generalizable.

We collected 15 million tweets based on 32 broad keywords relating to three types of risky behaviors associated with the transmission of HIV: drug use (e.g. meth), risky sexual behaviors (e.g. bareback), and other STIs (e.g. herpes). We then hand coded a subset of 2,537 unique tweets using a crowd-sourceable "game" that can be distributed online. This hand-coded set was used to train a simple n-gram classifier, which resulted in an algorithm to select relevant tweets from the larger database. We then generated geocodes from text locations provided by the tweet author, supplemented by the ~1% of tweets that are already geolocated. We scaled these geocodes to the state and county levels, which allowed us to compare HIV prevalence in our collected data with public health data.

Results

We present the correlation between behaviors identified in social media and the corresponding impacts on disease incidence across a large population. Hand coding revealed that 34% of tweets with one or more of the 32 initial keywords was relevant to behaviors associated with HIV transmission. Among the three categories of initial search terms, the drug category yielded 21% true positives, compared to 9% for risky behaviors, and 2% for other STIs. The n-gram classifier measured 66% sensitivity and 44% specificity on a test set. In addition, our geolocation algorithm found coordinates for 88% of text locations. Of those, a test sample of 59 text locations showed that 83% of geolocations are correctly identified. These components combine to form an analysis pipeline for detecting risky behaviors across the United States.

Conclusions

We present a surveillance methodology to help sift through the vast volumes of these data to detect behaviors and determinants of health contributing to both disease transmission and chronic illness. This effort allows for identification of at-risk communities and populations, which will facilitate targeted, primary and secondary-prevention efforts to improve public health.

Keywords

social media; hiv/aids; digital epidemiology

Acknowledgments

We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by: NIH MIDAS Grant 2U01GM070694-09 and DTRA CNIMS Contract HDTRA1-11-D-0016-0001

References

[1] Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., et al. (2012). Digital Epidemiology. (P. E. Bourne, Ed.)PLoS Computational Biology, 8(7), e1002616. doi:10.1371/journal.pcbi.1002616

[2] Achrekar, H., Lazarus, R., & Park, W. C. (2011). Predicting Flu Trends using Twitter Data. The First International Workshop on Cyber-Physical Networking Systems (pp. 713–718).

[3] Neill DB. Fast Bayesian scan statistics for multivariate event detection and visualization. Stat Med. 2011Feb.28;30(5):455–69.

*Caitlin Rivers
E-mail: cmrivers@vbi.vt.edu