Tracking Alcohol and Marijuana Usage and Behaviors from Social Media using Oak Ridge Bio-surveillance Toolkit

Arvind Ramanathan, Shannon P. Quinn, Laura L. Pullum, and Chad A. Steed

Abstract— **We present an overview of how the Oak Ridge Bio-surveillance Toolkit (ORBiT) can track and monitor alcohol and marijuana usage across the United States (US). ORBiT was developed as a platform to integrate various public health related data streams to support situational awareness of emerging public health concerns such as new infectious diseases. However, with the recent surge of social media data, we hypothesize that ORBiT can also be used to track and monitor the use of alcohol and marijuana amongst young people across the US. Our preliminary studies demonstrate that while it is possible to obtain overall trends in alcohol and marijuana use, it can be challenging to extract behavioral data from social media. We also highlight some of the emerging challenges in using social media datasets to extract meaningful information regarding alcohol/marijuana use.**

I. Introduction

Social influences are key in defining health behaviors; in particular, people with more friends/social contacts are happier and people that are highly influential (i.e., connected to many people) are more susceptible to the (health-related) benefits and risks of their connections [1]. While social influences have a positive impact on decreasing the number of young people starting to smoke and promoting active, healthy life-styles, peer pressure plays a significant role in encouraging young adults and adolescents to try alcohol and/or other drugs such as marijuana. With the explosive growth of social media, especially in the context of online user-generated/self-reported content from various social networking sites including Facebook, Twitter and Pinterest, and their ubiquitous reach toward young adults and adolescents, we hypothesize that the social media content can be used to track how young adults use alcohol and marijuana.

Recently, we developed Oak Ridge Bio-surveillance Toolkit (ORBiT) as a scalable computing platform to analyze public health related data streams to monitor and track emerging infectious disease outbreaks [2]. In particular, we have shown that ORBiT can analyze large volumes of electronic healthcare reimbursement claims datasets to extract multi-scale spatial and temporal patterns of influenza spread across the US [3]. Further, ORBiT was also used to quantify the co-occurrence of asthma and flu during the 2009-2010 pandemic flu season [4]. We also recently augmented ORBiT with sequential pattern mining tools to extract clinical trajectories and to quantify commonalities in treatment patterns across large patient cohorts [5].

*Research supported by ORNL Seed Project 7280 and Laboratory Director Research and Development Project 7417.

A. Ramanathan, L.L. Pullum and C.A. Steed are with the Computational Science and Engineering Division, Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (phone: 865-576-7266; e-mail: {ramanathana,pullumll,steedca}@ornl.gov).

S.P. Quinn is with the Department of Computer Science, University of Georgia, Athens, GA USA (e-mail: squinn@cs.uga.edu).

The techniques developed as part of ORBiT can also be applied to social media datasets to extract patterns of alcohol and marijuana usage. To monitor the use of alcohol and marijuana, we used Twitter as a primary source of self-reported data. Twitter is a popular microblogging site where posts typically capture small (within 140 characters) self-reported status messages (referred to as "tweets"). Using the Twitter streaming API, we obtained our data by filtering posts based on specific keywords representing alcohol/marijuana use from urban slang dictionaries. This allowed us to capture roughly 1,000 tweets every minute, representing a relatively large collection of over a million tweets even within a single week.

Using natural language processing tools, we were able to track how often users mention alcohol/marijuana (and related terms) in their tweets. Further, we were also able to extract temporal patterns in the usage of alcohol/marijuana, based on frequency of these terms being used. However, since we were not able to obtain access to a single user's history of tweets, it is difficult to extract patterns of individual behaviors observed from these tweets. In particular, it is difficult to discern whether individuals are progressing towards more addictive behaviors versus weaning off of alcohol/marijuana use. In order to obtain such behavioral signatures, we propose to augment mining of social media data streams with electronic healthcare reimbursement claims as well as targeted clinical studies that link self-reported substance use to Twitter activity. This integrated approach will enable us to determine the risks of individuals with adverse effects of alcohol/marijuana abuse. Further, it can lead to personalized intervention strategies that can also potentially be administered through social media campaigns.

References

[1] S. Galea, A. Nandi, D. Vlahov, The social epidemiology of substance abuse. *Epidemiologic Reviews* (2004), 26: 36-52.

[2] A. Ramanathan, L.L. Pullum, T.C. Hobson, C.A. Steed, S.P. Quinn, C.S. Chennubhotla, S. Valkova, ORBiT: Oak Ridge biosurveillance toolkit for public health dynamics. *BMC Bioinformatics* (2015), 16 (Suppl. 17).

[3] L.L. Pullum, A. Ramanathan, Oak Ridge Biosurveillance Toolkit: Scalable machine learning for public health surveillance, *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on (2014),* Orlando, FL.

[4] A. Ramanathan, L.L. Pullum, T.C. Hobson, C. G. Stahl, C.A. Steed, S.P. Quinn, C.S. Chennubhotla, S. Valkova, Discovering Multi-Scale Co-Occurrence Patterns of Asthma and Influenza with Oak Ridge Bio-Surveillance Toolkit. *Frontiers in public health* (2015), 3 (1): 182.

[5] K. Malhotra, T.C. Hobson, S. Valkova, A. Ramanathan, L.L. Pullum, Sequential Pattern Mining of Electronic Healthcare Reimbursement Claims: Experiences and Challenges in Uncovering How Patients are Treated by Physicians. *Workshop on Mining Big Data to Improve Clinical Effectiveness, Proceedings of the IEEE Big Data 2015*, Santa Clara, CA, USA.