OJPHI

ISDS 2014 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

ChatterGrabber: A Lightweight Easy to Use Social Media Surveillance Toolkit

James T. Schlitt*, Bryan Lewis and Stephen Eubank

Virginia Bioinformatics Institute, Virginia Polytechnic Institute, Blacksburg, VA, USA

Objective

To formally introduce ChatterGrabber, an open source, natural language processing based toolset for public health social media surveillance. ChatterGrabber is designed to collect and categorize a high volume of content at a low cost, providing a readily deployable solution for Epidemiologists to track emergent outbreaks in the field and a signal for syndromic surveillance.

Introduction

Despite numerous successes in using social media to detect food borne illness and to predict influenza trends, the use of social media as a public health tool has yet to gain widespread adoption. While social media data cannot directly diagnose illness, aggregate trends in symptom proliferation may readily be observed. Such trends may allow a health agency to watch for signs and symptoms related to target conditions within its jurisdiction. Further, social media surveillance offers a distinct advantage in immediacy and sensitivity as it is not dependent upon infected individuals seeking care for reportable illnesses and as such information is not delayed by the handling, transfer, and processing of reports. These advantages may enable the earlier preparation and initiation of scaled response sequences during public health emergencies. Such data may also yield additional evidence through shared symptoms, rumors, and observations crucial to an epidemiological investigation.

Methods

ChatterGrabber uses the Twitter RESTful API to perform high frequency searches across a local, regional, or international geographic area. After a sufficient volume of posts has been obtained via a keyword search, posts are manually scored and used to train a classifier via the Natural Language Tool Kit's guided machine learning. Classifiers and run parameters may be configured by an investigator via an online interface to select n-grams observed, feature frequency limits for training set inclusion, and machine learning methodologies. Supported machine learning methods include naive-bayes, maximum entropy, decision tree, and multi-category support vector machines.

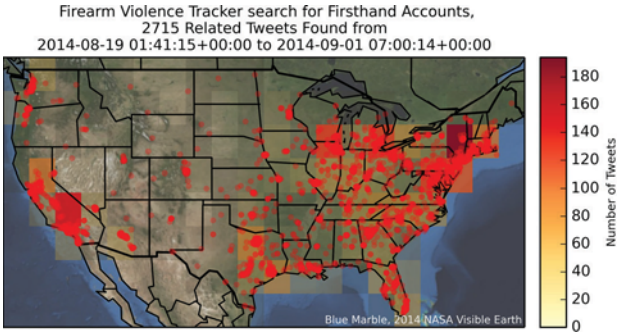
Weak classifiers are optimized by a distributed genetic algorithm in which random parameter sets are bred, run, and scored to select for the strongest. Classifiers are scored by the unweighted mean of the products of each category's sensitivity and specificity. This prevents classifiers from succeeding by over-applying to irrelevant posts or by classifying all posts as a common category. Once a strong classifier is obtained, posts are processed based upon the classifications chosen. All content is timestamped and located by the Google Maps API from attached coordinates or profile locations. Reports and visualizations are generated nightly and emailed to subscribers, providing an intuitive, distributable summary of area conditions at the start of each day.

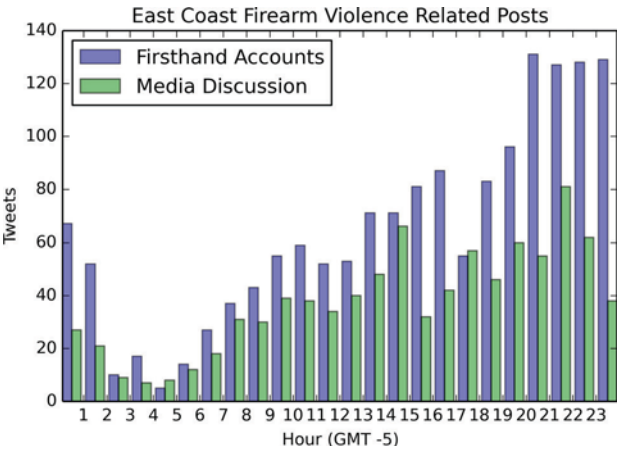
Results

The system has been successfully adapted to track Norovirus, tick-borne illness, firearm violence, vaccine sentiment, first responder emergencies, regional Ebola rumors, and Equine Herpes Virus for State, Federal, and Local health agencies. Output from ChatterGrabber has been used to help construct Ebola models, to monitor rural Norovirus epidemics, and incorporated into a health department dashboard system.

Conclusions

ChatterGrabber provides a cheap, effective, and readily deployable means for epidemiologists to track emergent outbreaks. The combination of natural language processing and geographic region directed searches allows one to surveil any user defined jurisdiction and hasten illness identification. The use of social media data allows for rapid identification of emerging outbreaks and provides a wealth of soft information, quantitative, and qualitative data to aid in an investigation.





Keywords

twitter; surveillance; dashboard; ChatterGrabber; natural language processing

Acknowledgments

Funding for ChatterGrabber was supported by the National Institutes of General Medical Sciences of the National Institutes of Health under award number 5U01GM070694-11.

*James T. Schlitt
E-mail: jschlitt@vbi.vt.edu