ISDS 2013 Conference Abstracts

Developing a Social Media System for Biosurveillance

Julie Waters*, Kristina Howard, Heather Baker and Caroline Brown

National Biosurveillance Integration Center (NBIC), Department of Homeland Security (DHS), Washington, DC, USA

Objective

Through ongoing and future projects we will examine the utility of social media data for biosurveillance, including machine learning approaches for algorithm development, as well as the system and organizational architectures required to implement an operational system.

Introduction

Much attention has been given recently to the purported ability of social media to provide early warning and/or situational awareness and event characterization during a biological event of national concern. The National Biosurveillance Integration Center's (NBIC) innovation project on Social Media Analysis seeks to demonstrate the viability of extracting relevant, health information from social media data, with the ultimate goal to establish an operational social media system for biological event surveillance.

Early work in this project has focused on demonstrating the relevance of social media to the biosurveillance problem through data analysis and algorithm development. Preliminary assessments of a commercial social media product also yielded valuable insights for the system architecture required to support such an operational tool. In addition to continued analysis of data utility (algorithm development) and system architecture, future work will include development of a comprehensive concept of operations (CONOPS) for implementation and use of a social media capability within the NBIC.

Methods

Through an ongoing algorithm development project with the Naval Surface Warfare Center, Dahlgren Division, NBIC is demonstrating and characterizing the utility of social media for early indication and situational awareness of significant biological events using Twitter data. Researchers at Dahlgren developed and refined a Twitter-specific ontology of health-related terms to assess the health of users. Using a training set of nearly 2000 tagged "sick" or "not sick" tweets, they applied several algorithms, including Support Vector Machines (SVM), Random Forest, K-Nearest Neighbor, and Latent Dirichlet Allocation (LDA) to determine the viability of using supervised machine learning (ML) to automatically classify health-related tweets and analyze trends over time and geography.

Additionally, we investigated the system architecture requirements of an operational system by examining the Social Media Analytics (SMA) toolkit, a COTS product by SAS Analytics. By tailoring this commercial product to detect trends around a representative health scenario we determined how well this existing architecture would meet the needs of a biosurveillance system.

Future efforts will include the development of a CONOPS for implementing and operationalizing a social media capability within NBIC. How will we use social media in the future, and how will it impact our current operating procedures and products? A thorough understanding of NBIC customer needs and how social media can meet them is essential. Just as necessary is a characterization of the uncertainty around these data and the analyses they support.

Results

Preliminary classification results using leave-one-out validation on the training set of tweets showed classification accuracy rates greater than 85% for most ML algorithms tested. These early results show it is possible, and practical, to collect and classify health-related Twitter data from a national, geo-located stream for trend analysis.

Through the SMA toolkit, we gained valuable insight for the architecture required to support an NBIC application of social media. Technical components of an operational system must include big data infrastructure; social media data sources; a content categorization engine to employ developed algorithms; a user portal; an analyst workbench; and a processing pipeline for new data. One of the biggest challenges for operational deployment appears to be scalability across the significant number of health-related query terms of relevance for biosurveillance across human, animal, plant, food, and environmental domains.

Conclusions

Social media is a unique category of data. Application of this data to the biosurveillance problem is yet largely unexplored. Initial analysis shows promising capacity for health information within social media. Significant investment and future effort is required in the areas of algorithm, system architecture, and CONOPS development to enact an operational social media system for biosurveillance.

Keywords

Department of Homeland Security (DHS); social media; machine learning; biosurveillance

*Julie Waters
E-mail: julie.waters@hq.dhs.gov