Social Media Analytics and Research Test-bed (SMART Dashboard)

Ming-Hsiang Tsou
mtsou@mail.sdsu.edu

Chin-Te Jung
chinte.jung@gmail.com

Chris Allen
allen14@rohan.sdu.edu

Jiue-An Yang
jyang@mail.sdsu.edu

Jean-Mark Gawron
gawron@mail.sdsu.edu

Brian H. Spitzberg
spitz@mail.sdsu.edu

Su Han
suhanmappingideas@gmail.com

Center for Human Dynamics in the Mobile Age, San Diego State University
5500 Campanile Drive, San Diego, CA 92182-4493
+1-619-5940205

ABSTRACT

We developed a social media analytics and research testbed (SMART) dashboard for monitoring Twitter messages and tracking the diffusion of information in different cities. SMART dashboard is an online geo-targeted search and analytics tool, including an automatic data processing procedure to help researchers to 1) search tweets in different cities; 2) filter noise (such as removing redundant retweets and using machine learning methods to improve precision); 3) analyze social media data from a spatiotemporal perspective, and 4) visualize social media data in various ways (such as weekly and monthly trends, top URLs, top retweets, top mentions, or top hashtags). By monitoring social messages in geo-targeted cities, we hope that SMART dashboard can assist researchers investigate and monitor various topics, such as flu outbreaks, drug abuse, and Ebola epidemics at the municipal level.

Categories and Subject Descriptors

Collaborative and Social Computing – Social Media, Collaborative and Social Computing design and evaluations – Social network analysis, Collaborative and social computing systems and tools— Social Networking Sites, Spatial-temporal systems— Geographic information systems

Keywords

Social Media Analytics, Twitter, Disease Outbreak, Geo-targeted, Spatiotemporal.

1. INTRODUCTION

Careful mining of social media messages (tweets and Facebook posts) can reflect the trends of human dynamics, such as where flu is spreading [1-2], when social movement ideology is diffusing [3], and how urban mobility patterns reveal geospatial and social functions. Despite the new and evolving nature of social media,

various models and theories are emerging to account for the ways in which the message contents and exchange dynamics both reflect and drive human processes in realspace [4-5]. Seeking to integrate several traditional (e.g., framing theory, narrative theory, diffusion of innovations, information theory, communicative competence theory), Spitzberg [6] proposed a multilevel model of meme diffusion (M^3D). Memes are any replicable unit of cultural transfer of information, so social media messages are all potential memes. The degree to which the diffusion of memes through social media networks can be modeled becomes a major theoretical challenge for businesses, governments, celebrities, and all those seeking to understand the human dynamic of communicative influence [7].

The M^3D model anticipates that certain features of memes or social media messages (e.g., novelty, repetition, etc.), of communicators (e.g., source credibility, network centrality, etc.), of structural network effects (e.g., network span, homophily, etc.), of subjective network structure (e.g., counter-memes, cascades, etc.), of societal processes (e.g., publicity campaigns, stage of diffusion, etc.), and of geo-technical factors (e.g., geospatial proximity, population density, etc.) will predict meme diffusion. In any given context, such processes may reveal unique diffusion maps or patterns across time and space, varying by rapidity of diffusion, exhaustion and duration of diffusion, and evolution of message variation. To the extent that such unique patterns can reveal distinct correspondence to realspace activities, they provide an important window into developing surveillance and intervention programs to serve the public interest in diffusing time and space sensitive information (e.g., disease diffusion and/or treatment, drug abuse or diffusion, natural disaster or crisis response). In other contexts, such surveillance may map important applications for organizations concerned with mapping or stimulating institutional reputation or product diffusion (e.g., academic reputation).

We developed a web-based social media analytics and research testbed (SMART) dashboard with geo-targeted social media application programming interfaces (APIs) (for Twitter) to provide real-time surveillance for various topics. SMART dashboard can be used to track multiple topics with various keywords, including disease outbreaks, drug abuse, regional wildfires in Southern California, and university-related activities and messages (URL: http://vision.sdsu.edu/hdma/smart).

SMART dashboard is built by using multiple data mining programs, GIS methods, and advanced geo-targeted social media

APIs to track selected topics in space and over time. Researchers can utilize SMART dashboard to visualize, characterize, and predict trends in these topics in different cities over time. Local healthcare providers, hospital staff, government officials, first responders, and other stakeholders can access the online, web-based dashboard without local desktop installation. The following features of SMART dashboard underscore its significance and innovativeness:

1. SMART dashboard captures and updates the spatial nature of social media messages every day and can thus evaluate patterns of messages in diverse cities and track patterns of diffusion geographically.

2. The front-end of SMART dashboard can display the dynamic temporal trends of social media messages (daily, weekly and monthly) with interactive selection tools (in [Trend] toolbox).

3. The back-end of SMART dashboard includes data filtering and machine learning components to remove noise and errors in different subjects. These procedures can facilitate more accurate analysis and tracking of disease outbreaks and drug abuse.

4. Users can select a [Target City] to analyze the temporal trend and top messages from individual cities or use the default view to display the aggregated trend analysis by combining messages from all cities in the list.

5. The Geo-targeted APIs and the tweets map displayed in the online interface can help users understand the differences of social media messages between cities and provide valuable information for local health agencies and hospitals to track disease outbreaks or drug abuse in various cities.

The development of SMART dashboard initially focuses on five topics (Flu, Whooping Cough, Wildfire, Drugs, and Aztecs) because they represent a diverse range of patterns of use, targeted users, required analysis functions, and social values. Communication in general, and social media in particular, has been identified as a key element in each type of context: flu [8, 9], whooping cough [10], wildfire [11, 13], drugs [14, 15], and institutional affiliation and reputation [16, 18]. As anticipated by the M^{3}D model and numerous other theories, there is likely to be some degree of reciprocal influence and representation between social media communication and the activities that involve that communication. Mapping such media content and dynamics is the first step in revealing the degree of correspondence in socially relevant contexts of human activity.

Each topic collects a combination of keywords to be used in the Twitter Search APIs. Although our current SMART dashboard only focuses on the five topics, it can be extended to various subjects by using different keyword combinations and selected cities. The following is the list of keywords in each topic and the selected cities:

Flu: "flu" or "influenza" in top 30 (population) U.S. cities.
Whooping Cough: "whooping cough" or "pertussis" in top 30 U.S. cities.
Wildfire: "wildfire", "#4SFire", "#SanDiegoFire" in San Diego and Los Angeles. Recently, we increased the search city list to all top 30 U.S. cities.
Drugs: "salvia", "kush", "adderall", and "heroin" in top 100 U.S. cities.
Aztecs (the nickname of San Diego State University): "sdsu" or "#GoAztecs" or "#aztecs" in top 100 U.S. cities.

Based on our previous experiences, we found that the keywords for Twitter Search APIs are not case sensitive. A keyword search will collect both hashtag (#) keywords and user name (@). For example, search for "sdsu" keyword will also retrieve tweets containing "#sdsu" and "@sdsu". However, Twitter seems modified their APIs functions after November, 2014. Some of the rules of keywords might be changed.

The Twitter Search APIs used in SMART dashboard can retrieve tweets back to six to nine days before, within the circle area. This retrospect search function is useful for monitoring some unexpected events, such as earthquakes. On the other hand, the Twitter Streaming APIs can provide a live stream of new tweets (no historical tweets) by defining keywords or a region (within a bounding box). If users define a region using Streaming APIs, the search results will only include GPS-tagged tweets. SMART dashboard only used Twitter Search APIs.

2. BACKGROUND AND CONTRIBUTION

The development of SMART is inspired by the rise of the web observatory as a tool that can aid big data research, such as Google Flu Trends (http://www.google.org/flutrends/) and Health Map (http://www.healthmap.org/en/). Although social media and other forms of big data present exciting new opportunities for researchers to examine human behavior, there are many obstacles to obtaining, processing, and analyzing this information. Consequently, many groups have devoted their efforts to developing systems that facilitate the acquisition and summarization of such data [19, 22].

However, unlike other similar tools, SMART strongly emphasizes the geospatial characteristics of social media data. By giving users the ability to target topics in specific geographical regions and analyze the spatiotemporal patterns of social media messages, SMART fills a significant gap in the landscape of social media analytic systems.

3. SYSTEM DESIGN

SMART Dashboard has multiple components to search, process, and visualize social media messages from Twitter Search APIs. Figure 1 illustrates the system design of SMART dashboard.
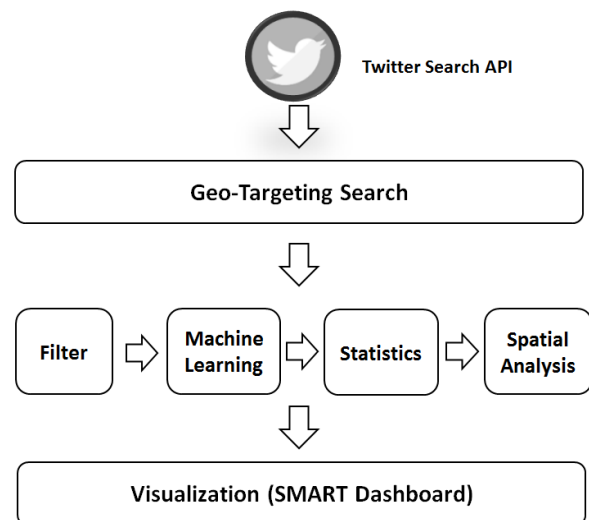

Figure 1. The system design of SMART dashboard.

We first developed geo-targeted search APIs, which can collect two types of spatial information from tweets: 1. geo-tagged locations provided by GPS-enabled devices, 2. self-reported locations specified in user profiles. Geo-tagged locations are latitude and longitude pairs created by mobile devices with built-in GPS receivers or by other geo-location features. Social media users specify the self-reported location and users can change it at any time. In the future, we will investigate the use of location inference methods, such as the study by Priedhorsky et al. [23], to enhance our geo-targeted search.

Since our SMART dashboard focuses on the city-level comparison and analysis of social media messages. We developed a geo-targeted method for normalizing tweeting numbers in each city based on their population size. Traditional city population size is based on the administrative boundaries and census tracks, which do not match the Twitter Search APIs method. We used GIS software to calculate the estimated population for each circle based on 2010 census tracts from the center of downtown to the radius defined by the Twitter Search APIs. Each of these city point buffers was joined with census tract centroids to determine which tracts should be included in our population calculations (Figure 2). Using the fine-grained census data allows us to gain a more accurate estimation of population, which greatly improves our ability to accurately normalize tweet counts for individual cities. Most of our city search methods are using a 17-mile radius buffer to cover major metropolitan areas in U.S. cities. Some cities are using larger radius or smaller radius, such as Phoenix (40 miles) or Anaheim/Irvine (10 miles).
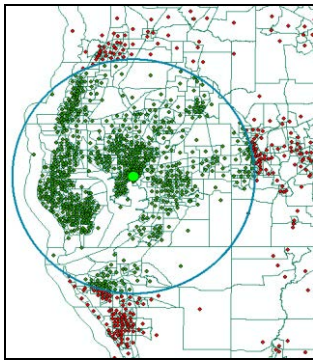

Figure 2. Using the geo-targeted APIs and GIS methods to collect tweets and calculate population from U.S. Cities.

Considering the huge datasets and query performance, SMART dashboard took the advantage of MongoDB, an open-source and NoSQL database, to store the search results. MongoDB is one of the most popular databases for Big Data which can provide high performance queries and supports spatial query from a very large datasets.

Data filtering and cleaning procedures are very important for social media analytics due to the frequent presence of noise. We developed comprehensive data filtering procedures (Figure 3) based on our previous research [24, 25]. The first step is to exclude retweets and tweets containing URL links, as our previous research has shown that these messages are often irrelevant. We also developed a machine learning classification procedure for filtering noise from our flu tweets. The goal of this procedure was to identify tweets that do not appear to indicate real-world cases of influenza so that they can be omitted from the statistical analysis. For classification, a linear support vector

machine (SVM) [26] was used, as this algorithm has shown to be effective for supervised learning tasks involving text. Twitter messages are transformed to numerical values using a term frequency – inverse document frequency (TF-IDF) model [27]. To train the SVM, 1,500 randomly sampled tweets from the 2012-2013 flu season containing the keyword "flu" were used to train the SVM classifier. Each of these tweets was manually inspected and classified as valid or invalid according to the likelihood that the message indicated an actual case of influenza, and this labeled data was used to train the SVM. We also used descriptive statistic methods to calculate the most popular retweets, the most popular URL (web pages), the most popular hashtags, and the most popular mentions (users as opinion leaders). Finally, we are developing additional components to analyze the hot spots of GPS-tagged tweets and to overlay with other GIS layers. The spatial analysis function was created in a separated viewer, called "GeoViewer" (http://vision.sdsu.edu/hdma/wildfire/).
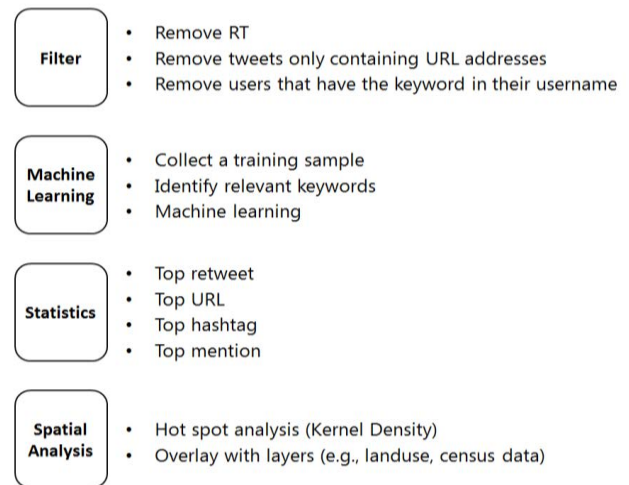
Figure 3. Four major data process procedures in SMART dashboard (filter, machine learning, statistics, and spatial analysis)

4. USER INTERFACE DESIGN

Different from traditional web GIS maps, the design of SMART dashboard needs to consider the interactive display of multimedia content (pictures, videos, text messages, hashtags) and maps together (Figure 4). The web-based user interface is built by using free open source programming libraries, jQuery and Leaflet APIs, to visualize and query tweets from a server-side database.

The following key features are included in the SMART dashboard to provide interactive query and visualization functions:

1. **Top index numbers** shows the number of tweets collected from one day, one week, or one month.

2. **Left panel** provides the list of targeted cities and the short cuts to different functions (Word Cloud [28], Trend, Top URL, Top Media, Top Cities, etc.)

3. **Trend Function** can provide interactive queries of actual tweeting texts by clicking on the point on the line chart. Users can switch the view to Daily, Weekly and Monthly.

4. **Word Cloud** function can show the most prominent conversation keywords in tweets within one day, past seven days, past 30 days, or combined all.

5. **Tweets in Cities** function can show the normalized tweeting rates (by city population) in each city using graduated cartographic symbols.

6. The dashboard also provides the top 10 list of top URLs (web pages), Hashtags (subjects), Retweets (forwarded messages), Mentions (opinion leaders), and Media (Pictures).
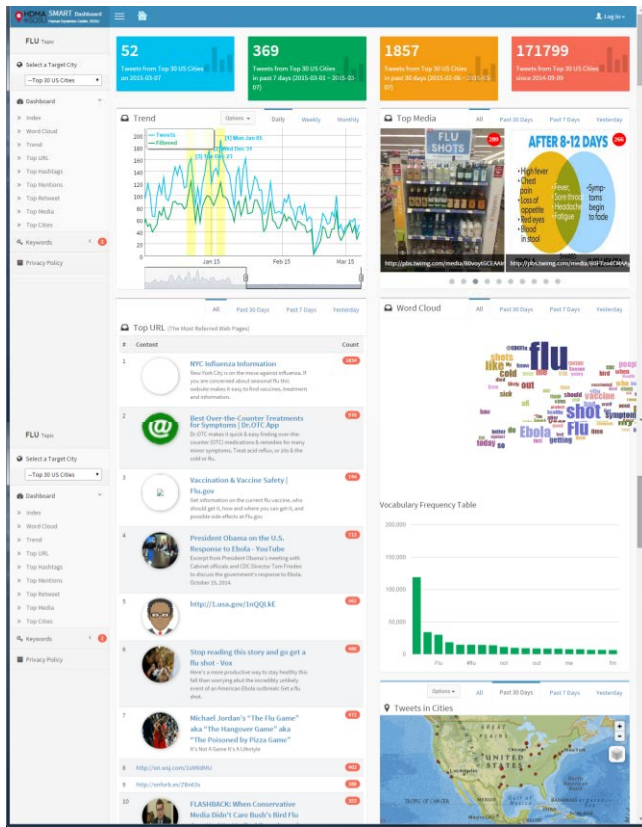

Figure 4. The user interface design in SMART dashboard

5. CASE STUDIES

This section introduces three case studies using SMART dashboard to monitor flu outbreaks in U.S. cities, track drug abuse trends in the U.S., and listen to messages related to Ebola outbreaks in West Africa, three U.S. Cities, and five U.S. airports. All three examples are related to public health issues in epidemiology. Surveillance is a key component in disease detection and public health practice, but traditional methods of collecting patient data and reporting to health officials are costly and time consuming. Social media analytics have a great potential to provide real-time monitoring and surveillance functions for disease outbreaks. We chose the three case studies to highlight the advantages of geo-targeted analysis and disease surveillance at the municipal level.

5.1 Flu Outbreaks in U.S. Cities

Figure 5 illustrates the weekly trend analysis of flu tweets in the SMART dashboard from the largest 30 U.S. cities from September 2014 to March 2015. We can use this graph to compare the official CDC regional ILI (Influenza-Like Illness) rates (the bottom graph in Figure 6). The SMART flu tweeting trend (top) is very similar to the CDC ILI trend (bottom) in FluView (www.cdc.gov/flu/weekly/fluviewinteractive.htm). Our previous analysis results indicated that correlation coefficients

(r-values) between weekly aggregated flu tweeting rates and disease occurrence (CDC regional ILI) were very high in many cities [2]. This interactive SMART dashboard can help health practitioners monitor and visualize daily changes of flu trends and related flu news. The daily monitoring capability in the SMART dashboard has great potential for local and state public health agencies and practitioners to integrate real-time information to investigate large-scale disease outbreaks.
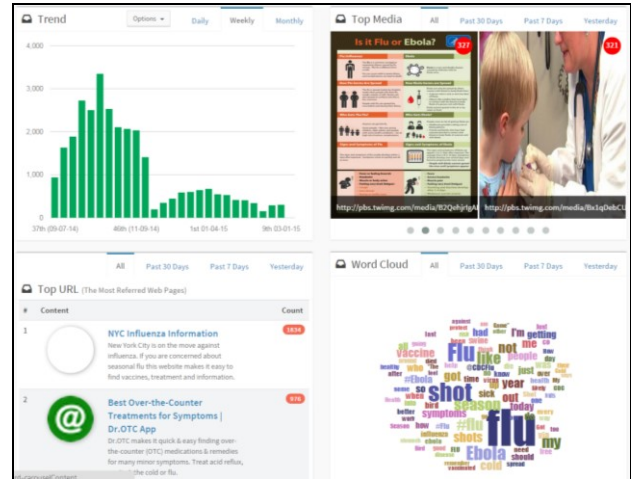

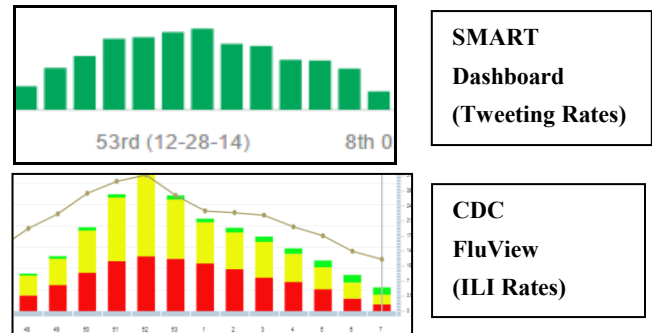Figure 5. Use SMART dashboard to analyze weekly flu tweeting rates for monitoring flu outbreaks.

Figure 6. Comparing the weekly results from SMART dashboard and the CDC FluView (ILI rates)

5.2 Drug Abuse and Misuse in U.S. Cities

Another case study of SMART dashboard is drug abuse and misuse in U.S. cities. Illicit drug use is a significant public health problem, ranking among the 10 of leading preventable risk factors for mortality and morbidity in developed countries [29]. In the U.S., an estimated 23.9 million adults were current illicit drug users in 2012 [30]. Surveillance is a key component among strategies to understand and address illicit drug use. Recent work has begun to explore promising non-obtrusive, real-time Big Data techniques and infoveillance tools [31] to obtain epidemiological information about drug use [32, 33]. However, these new techniques have not sufficiently captured the temporal and spatial nature of drug use trends. We hope that SMART dashboard can be used to analyze social media content from Twitter to characterize drug use trends in diverse geographic areas.

Figure 7 illustrates a high peak of drug related messages on 2014/10/07 due to a number of breaking news stories about a four

year-old girl carrying a bag of heroin packets in a local day care center in Delaware. The map view on the left side indicates the number of tweeting rate in top 100 U.S. cities (normalized by city population). A bigger circle indicates a higher number of tweeting rates in the selected city.
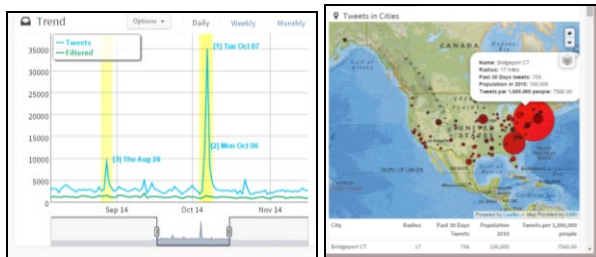
Figure 7. Drug abuse case study in SMART for identifying temporal trends and geographic distribution.

5.3 Tracking Ebola in West Africa and U.S. Cities

The recent emergence of the lethal Ebola virus in the U.S. and the West Africa caused the public significant fear and concern regarding the risk of an Ebola epidemic. Public health policies aim to minimize the impact of disease outbreaks, yet heightened fear and anxiety can often drive population behavior. An effective intervention must anticipate how populations will respond. Although our current health system and the Centre for Disease Control and Prevention (CDC) have good tools for reporting, monitoring, and measuring the spread of disease infection cases, we do not have effective tools to measure public perception, fear and response during significant public health crises. Such panics can be counterproductive, diverting needed resources and attention away from more effective interventions.
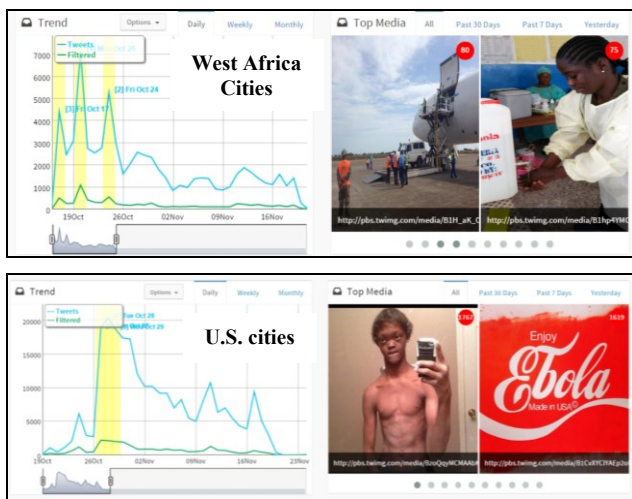
Figure 8. Monitor the Ebola outbreaks in West Africa and U.S. cities.

One potential use of the SMART dashboard is to monitor public opinions and responses during or after disease outbreaks or disaster events. Figure 8 illustrates the most popular media (pictures) sharing in Twitter between three U.S. cities (New York, Dallas, and Cleveland) and 12 West Africa cities. The temporal trends of Ebola related tweets in U.S. cities are very different from

the 12 West Africa cities. The most popular pictures shared in Twitter in U.S. cities are very negative or sarcastic (such as "Enjoy Ebola"). On the other hand, the most popular pictures and messages shared in West Africa cities are mainly disease prevention-oriented information or medical aid from international organizations. This case study demonstrates that SMART dashboards can be used to listening to the public opinions during crisis events or disease outbreaks.

6. CONCLUSION

Real time public health information capture using social media is now at the forefront of behavioral measurement, disease surveillance, and health promotion in many public health research activities. We developed SMART dashboard to track spatiotemporal trends in multiple case studies, including flu outbreaks, drug abuse, and Ebola epidemics. Healthcare providers and public health officials may be able to use SMART dashboard to allocate medical resources in an efficient and effective way.

Our team approaches the issues described above in a unique manner by analyzing social media messages from a spatiotemporal perspective. Different cities and regions may reveal different patterns of social media messages and trends. By analyzing the context of social media messages (linking place and time together), we can discovery more meaningful patterns and insights of disease outbreaks and social activities.

The development of SMART dashboard also has some challenges. Privacy is a main concern for social media analytics tools. Our team is aware of this concern and we try our best to protect the privacy of social media users. With respect to Twitter messages, we only collect public tweets available from the public Twitter APIs. In the SMART dashboard, we also expressed our Privacy Policy explicitly as "*If you have any concerns about the privacy issues in our web applications, please Email us (hdmasdsu@gmail.com). After verify your information, we will remove specific social media contents based on your requests.*" Another possible privacy protection method is to convert all users to an anonymous ID. However, the anonymization method may prevent the future analysis of social networks in these messages. We need to find a balance between the protection of user privacy and the usefulness of social media messages.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Nagel, A.C., Tsou, M.H., Spitzberg, B.H., An Li, Gawron, J.M., Gupta, D.K., Yang, J.A., Han Su, Peddecord, K.M., Lindsay, S., Sawyer, M.H. 2013 The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using tweets. *The Journal of Medical Internet Research* 15, 10, 237. doi:10.2196/jmir.2705.

[2] Aslam, A.A., Tsou, M.H., Spitzberg, B.H., An L, Gawron, J.M., Gupta, D.K., Peddecord, K.M., Nagel, A.C., Allen, C., Yang, J.A., and Lindsay, S. 2014. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research* 16, 11. doi:10.2196/jmir.3532.

[3] Ming-Hsiang, T., Yang, J.A., Lusher, D., Han, S., Spitzberg, B., Gawron, J.M., Gupta, D., and An, L.2012. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in 2012 U.S. presidential election. *Cartography and Geographic Information Science*.

[4] Cheung, C.M., and Thadani, D.R. 2012. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems* 54, 1, 461-470.

[5] Wang, C., and Zhang, P. 2012. The evolution of social commerce: The people, management, technology, and information dimensions. *Communications of the Association for Information Systems* 31, 5, 1-23.

[6] Spitzberg, B.H. 2014. Toward a model of meme diffusion (M³D). *Communication Theory* 24, 3, 311-339. doi:10.1111/comt.12042

[7] Kamel Boulos, M.N., Resch, B., Crowley, D.N., Breslin, J.G., Sohn, G., Burtner, R., and Chuang, K.S. 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International Journal of Health Geographics* 1067. doi:10.1186/1476-072x-10-67

[8] Culotta, A. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources & Evaluation* 47, 1, 217-238. doi:10.1007/s10579-012-9185-0

[9] Nagar, R., Yuan, Q., Freifeld, C.C., Santillana, M., Nojima, A., Chunara, R., and Brownstein, J.S. 2014. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research* 16, 10. doi:10.2196/jmir.3416

[10] Olsen, J. 2013. Infodemiology to improve public health situational awareness: An investigation of 2010 Pertussis outbreaks in California, Michigan and Ohio. Ph.D. dissertation, Gillings School of Global Public Health, University of North Carolina at Chapel Hill. http://gradworks.umi.com/3562785.pdf

[11] Helsloot, I., and Groenendaal, J. 2013. Twitter: Underutilized potential during sudden crisis? *Journal of Contingencies and Crisis Management* 21, 3, 178-183. doi: 10.1111/1468-5973.12023

[12] Slavkovikj, V., Verstockt, S., Van Hoecke, S., and Van de Walle, R. 2014. Review of wildfire detection using social media. *Fire Safety Journal* 68, 109-118. doi:10.1016/j.firesaf.2014.05.021

[13] Sutton, J., Spiro, E. S., Johnson, B., Fitzhugh, S., Gibson, B., and Butts, C. T. 2014. Warning tweets: serial transmission of messages during the warning phase of a disaster event. *Information, Communication & Society* 17, 6, 765-787.

[14] Hanson, C. L., Burton, S. H., Giraud-Carrier, C., West, J. H., Barnes, M. D., and Hansen, B. 2013. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research* 15, 4, e62. doi:10.2196/jmir.2503

[15] Young, S. D., and Shoptaw, S. 2013. Stimulant use among African American and Latino MSM social networking users. *Journal of Addictive Diseases* 32, 1, 39-45. doi:10.1080/10550887.2012.759859

[16] Jianhong, M., Junwen, F., Liuliu, Z., and Ziran, X. 2014. The construction and application research of crisis early warning mechanism of reputation of colleges and universities in the new media environment. *Journal of Chemical & Pharmaceutical Research* 6, 6, 202-209.

[17] Sagun, K. K. 2013. Internet memes as an information dissemination tool for libraries: The Ateneo de Manila University experience. *Procedia—Social and Behavioral Sciences* 103, 542-550. doi: 10.1016/j.sbspro.2013.10.371

[18] Snoeijers, E. M., Poels, K., and Nicolay, C. 2014. #universitycrisis: The impact of social media type, source, and information on student responses toward a university crisis. *Social Science Computer Review* 32, 5, 647-661. doi:10.1177/0894439314525025

[19] McKelvey, K., and Menczer, F. 2013. Design and prototyping of a social media observatory. In *Proceedings of the 22nd international conference on World Wide Web companion*.. International World Wide Web Conferences Steering Committee. 1351-1358.

[20] Gruzd, A., Haythornthwaite, C., Paulin, D., Absar, R., and Huggett, M. (2014, March). Learning analytics for the social media age. In *Proceedins of the Fourth International Conference on Learning Analytics and Knowledge*.. ACM. 254-256.

[21] Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. 2009. Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine* 360, 21, 2153-2157.

[22] Boulos, M. N. K., Sanfilippo, A. P., Corley, C. D., and Wheeler, S. 2010. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine* 100, 1, 16-23.

[23] Priedhorsky, R., Culotta, A., and Del Valle, S. Y. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* ACM.1523-1536.

[24] Tsou, M. H. and Leitner, M. 2013. Editorial: Visualization of Social Media: Seeing a Mirage or a Message? In Special Content Issue*:* Mapping Cyberspace and Social Media. *Cartography and Geographic Information Science* 40,2,55-60.

[25] Tsou, M.H., Kim, I.H., Wandersee S., Lusher D., Li A.,Spitzberg, B., Gupta D., Gawron J.,Smith J. , Yang J.A., Han, S.Y. 2013. Mapping Ideas from Cyberspace to Realspace: Visualizing the Spatial Context of Keywords

from Web Page Search Results. *International Journal of Digital Earth*, 316-335.

[26] Salton, G, and Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24, 5, 513–23.

[27] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning (ECML), Springer.

[28] Kuo, Byron YL, et al. 2007. Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web*. ACM.

[29] U.S. Preventive Services Task Force (USPSTF). Screening for Illicit Drug Use: U.S. Preventive Services Task Force Recommendation Statement. January 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/druguse/drugrs.htm

[30] U.S. Department of Health and Human Services (USDHHS). (2013). Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings. NSDUH SeriesH-46, HHS Publication No. (SMA) 13-4795. Rockville, MD: Substance Abuse and Mental Health Services Administration. Retrieved from http://www.samhsa.gov/data/NSDUH/2012SummNatFindDetTables/NationalFindings/NSDUHresults2012.htm

[31] Eysenbach, G. 2009. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research* 11, 1, e11.

[32] Chary, M., Genes, N., McKenzie, A., and Manini, A.F. 2013. Leveraging social networks for toxicovigilance. Journal of Medical Toxicology 9,2, 184-191.

[33] Cameron, D., Smith, G. A., Daniulaityte, R., Sheth, A. P., Dave, D., et al. 2013. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. Journal of Biomedical Informatics, 46, 985-997. Retrieved from http://dx.doi.org/10.1016/j.jbi.2013.07.007.