Social Media: A Systematic Review to Understand the Evidence and Application in Infodemiology

Stacey Guy, Alexandria Ratzki-Leewing, Raphael Bahati,
and Femida Gwadry-Sridhar

Lawson Health Research Institute,
Commissioners Rd E. 801, N6C 5J1 London, Canada
stacey.guy@sjhc.london.on.ca
{alexandria.ratzkileewing,raphael.bahati}@lawsonresearch.com
femida.gwadry-sridhar@lhsc.on.ca

Abstract. Social media represents a new frontier in disease surveillance. Infoveillance allows for the real-time retrieval of internet data. Our objective was to systematically review the literature utilizing social media as a source for disease prediction and surveillance. A review of English-language conference proceedings and journal articles from 1999 to 2011 using EMBASE and PubMed was conducted. A total of 12 full-text articles were included. Results of these studies show the use of open-source micro-blogging sites to inform influenza-like-illness monitoring. These results inform recommendations for future research directions.

Keywords: social media, review, population surveillance, data mining.

1 Introduction

Today, the number of social media users continues to skyrocket with rates of participation on social networking sites already quadrupling from 2005-2009 [1]. As an easily accessible, highly cost-effective and interoperable system, social media opens doors to a better understanding of community creation, providing fast access to information anywhere in the world, 24 hours a day [2,3].

Due to the popularity of online communication, open-source social media platforms present excellent opportunities in health research [4]. Using a strategy called infoveillance, real-time online data can be systematically mined, aggregated and analyzed to inform public health and policy [5,6]. More specifically, social media can be used as a relevant and real-time source of epidemic intelligence [4].

Mining online information can provide insight to abnormal patterns of disease and aid in predicting disease outbreaks. Various studies have confirmed the potential of infoveillance to advance epidemic intelligence.

The aim of this paper is to illustrate how data generated through social media can be used to inform planning and implementation of strategies to address communicable disease emergence - in turn, changing the future of health research.

P. Kostkova, M. Szomszor, and D. Fowler (Eds.): eHealth 2011, LNICST 91, pp. 1–8, 2012.

This paper is organized as follows. The methodology used in this systematic review is described in section 2. The results of the review are presented in section 3. A discussion of these findings and their implications are debated in section 4.

2 Method

2.1 Data Sources

We conducted a systematic review of the literature utilizing the bibliographic databases EMBASE and PubMed in June 2011. The following keywords were used to search EMBASE; keywords were divided into three categories: (1) *Disease (Early detect$, Pandemic$, Epidemic$, Communicable disease$, Early diagnosis)*, (2) *Medium (Information technol$, Internet, Mass medium, Medical computing, Social Media, Social network$, Geolocation)*, and (3) *Methodology (Disease surveillance, Monitor$, Disease control, Algorithm$, Data min$, Query process$, Information retriev$).* MeSH terms used to conduct the search in PubMed were also divided into three categories: (1) *Disease (Pandemics, Communicable diseases, Disease outbreaks, Early diagnosis)*, (2) *Medium (Communications media, Databases, factual, Internet*, User-computer interface*)*, and (3) *Methodology (Population surveillance, Information storage and retrieval, Forecasting, Data mining, Sentinel surveillance).* The keyword in which the article was indexed, the title of the article, and the abstract were searched for these categories described above. Unfortunately, terms such as 'infoveillance' have not yet been coded in these databases.

Search terms were chosen to reflect our objective – to review published research on disease surveillance using open-source social media. Conducting a generalized search in Google, elicited a few relevant publications to inform our search term selection. Index terms of these publications informed our formalized search strategy.

EMBASE and PubMed were chosen as they limit our search to published material in the arena of health care as opposed to using a computer-focused bibliographic engine that would require more terms to narrow the search to health. By including both databases we were able to search both North American and European published literature. The search was limited to both journal articles and conference proceedings – specifically, the terms 'conference paper' and 'proceeding' was used for EMBASE, while 'clinical conference' was used for PubMed) written in the English-language and published between 1999 and 2011 (for EMBASE, 'current' was chosen; for PubMed we used 1999/01/01 to 2011/12/01). The time period chosen reflects the addition of the term 'internet' to the PubMed index – no such MeSH term exists for social media however the term 'internet' is broad and encompassing.

2.2 Data Extraction and Synthesis

Using the search terms described above, a total of 384 journal articles and 1 conference proceeding was retrieved form PubMed and a total of 484 journal articles from EMBASE (with limits applied). In addition, our hand search (informed by the

references within the retrieved publications and Google searching) revealed 16 possible publications. The combined databases of 885 publications were searched for duplicates which resulted in a total of 287 publications being eliminated. Three of the authors (R.B, S.G, A.R-L) reviewed the publication abstracts to further establish relevance. Publications without abstracts were excluded. Publications that mined RSS feeds, survey data, physician records, medical records and search engines were excluded. Twelve publications focusing on data mining social media were reviewed.

3 Results

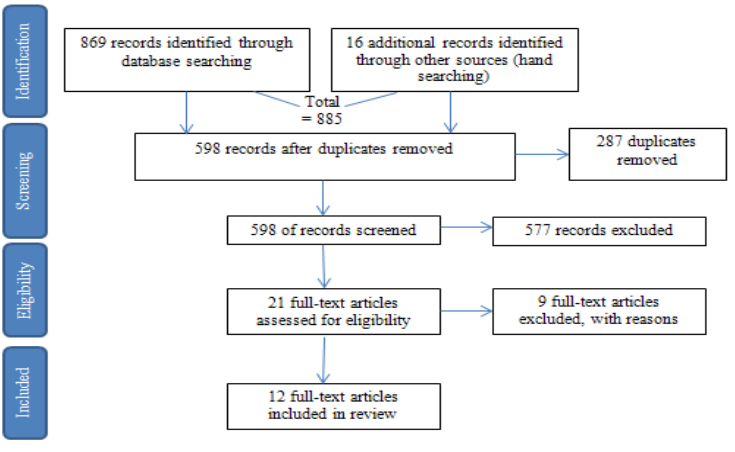

Fig. 1. This flow diagram illustrates the study selection process. A total of 885 publications were identified by bibliographic and hand searching. Through application of inclusion and exclusion criteria, 12 full-text publications were included in this review. This diagram is based upon The PRISMA Statement [7].

3.1 Study Characteristics

All publications reported focused on mining social media for the purpose of disease surveillance and prediction. As infodemiology is a relatively novel and emerging field, most studies were exploratory in design. The diseases of choice include influenza-like-illness (ILI) and H1N1.

3.2 Results of Individual Studies

Social Network Enabled Flu Trends (SNEFT) [8,9] is used to track and predict ILI activity. Tweets were retrieved (including metadata) with ILI content over approximately a 12 month period. SNEFT consists of separate data repositories where aggregated anonymous data is stored, and uses the ARMA model to predict ILI incidence. The web crawler was developed using Twitter's Search API. Regressive models were built and

evaluated with CDC (Centres for Disease Control) data. The system retrieved 4.7 million – retweets were extracted. To validate the observed trends, this data was compared to CDC data. The dataset containing no retweets and no tweets from the same user, resulted in the highest correlation (r= 0.9846) with CDC data.

Corley et al., [10] evaluated blog posts containing ILI keywords. Flu-related posts were extracted from 44 million posts collected over a 3 month period. A seven day period in posting was identified and verified. Categorized baseline trends were compared to CDC data to identify anomalies. Results show a significant correlation between the frequency of ILI posts per week and CDC data. In addition, Corley et al {{4750 Corley 2010; 4028 Corley 2010;}} collected 97,955,349 weblogs, micro-blogs and social media items pertaining to ILI data over a 20 week period. English language items containing relevant keywords were retrieved and grouped by month, week and day of the week. Flu-related data was compared to CDC data, and found to be highly correlated (r = 0.626 at 95% confidence).

Culotta et al., [4] have developed regression models using 574, 643 tweets collected over a 10 week period to predict ILI. To obtain a random sample of tweets the authors searched for common words. The percentage of tweets that reported ILI was estimated. Their findings indicated multiple regression out performs simple regression, keywords selected based on residual sum of squares is more effective than selecting keywords based on a correlation coefficient, and the best model of prediction (r=0.78) was one where "a simple bag-of-words classifier trained on roughly 200 documents can effectively filter erroneous document matches".

de Quincey & Kostkova [13] conducted a study to identify ILI trends present in tweets. The Twitter Search API was used to retrieve 100 tweets (including metadata). A PHP code parsed returned tweets (every minute) which were then saved to a MYSQL database. The system ran for 1 week in May 2009. A total of 135, 438 tweets containing 'flu' terminology were retrieved. The content was analyzed for trends using 'flu' in conjunction with other keywords ('swine' + 'flu'). Future plans involve the use of collocation analysis to identify trends.

Infovigil is an open-source infoveillance system which mines, analyzes and visually represents textual health-related data from Twitter [6,14]. Infovigil was used to plot term prevalence, and provide content analysis of tweets pertaining to H1N1. Two million tweets were retrieved over an 8 month period. English-language tweets were selected for, and retweets were excluded. Tweet patterns were influenced by media with the most commonly tweeted material being news (52.6%) Original tweets, as opposed to retweets, contained more personal experiences. There was very little misinformation found in the tweets (4.5%). The majority of automated queries correlated with manual coding results.

Lampos & Cristianini [15] developed a monitoring tool to track ILI patterns using UK specific Twitter data. Tweets containing symptom-related keywords were collected over 6 months during 2009. A daily average of 160,000 tweets were retrieved. This data, converted to a flu-score, was compared to weekly H1N1 reports from the Health Protection Agency. This resulting score correlated highly with reports (>95%). This method works independently of language, can determine self-diagnostic statements in tweets, and uses time series geolocated data.

Data Collector [16] is a system that uses social media as a source for real-time data. The backend consists of a web crawler, written in PHP and utilizing the Twitter Search API. Tweets are stored in a relational database according to UML class with 2 main categories: disease and location. Data Collector supplies a RESTful API divided into location, disease, and occurrence, which specifies methods and parameters through which one can access the database. The frontend web interface uses AJAX to produce real- time-graphs and maps. A dataset containing H1N1 tweets was collected from 6 European countries between May and July 2009. This system collected an average of 3200 tweets per day; 700 pertained to H1N1.

Signorini et al., [17] tracked sentiment and H1N1 activity using tweets. Keywords were used to retrieve tweets from the US. Prediction models were trained using CDC ILI values. Results are divided into 2 sets of data. The first saw 951,697 tweets from 334,840,972 retrieved over 34 days and was used to plot spikes of public interest. The second set represented over 4 million tweets from 8 million over a 3 month period. When analyzed, this showed no sustained interest in vaccine-related issues by the public. ILI estimates were gathered using a model trained on 1 million ILI tweets for 8 months. This model produced estimates that the authors believed were fairly accurate (average error = 0.28%; SD 0.23%). In order to garner real time estimates of ILI according to region, Signorini et al., [17] developed a model with region readings fitted to geolocated tweets. This model was less precise (average error = 0.37%) than the national weekly model.

The majority of articles included in the systematic review collected and examined data from Twitter (n=7); only 1 article looked at weblogs, micro-blogs, and social media. Studies demonstrated that a general correlation exists between ILI content in Tweets and CDC data. Research also indicated that Tweet patterns are strongly influenced by media, with news being the most commonly tweeted material. Finally, additional research is needed to determine the effectiveness of geo-location in garnering real-time estimates of ILI according to region.

4 Discussion

In realizing the potential of infodemiology in healthcare, it is important to consider the advantages and disadvantages of mining social media. Additionally, researchers must acknowledge and identify key target audiences; indeed, social media has specific target audiences with unique engagement behaviour specific to a platform that may or may not be representative of the population at large. Assessing the uses and potential of infodemiology in healthcare can improve user interaction, information access, evidence-based medicine, and knowledge representation.

4.1 Advantages and Disadvantages of Mining Social Media

Infoveillance can provide real-time, immediate and relevant information [6,14]. This is particularly useful when seeking timely and reliable data on the spread or severity of influenza [18]. Analyzing and disseminating real-time information can also

improve public access to health surveillance information. As data-mining sources utilize open-source information, the operating costs of these systems can be extremely low [6,14]. In addition 'mashups' are the new multi-taskers, capable of mining, categorizing, filtering, and visualizing online, real-time data on epidemics [19].

Certain pitfalls to mining social media exist. First, textual data can be difficult to classify and interpret since harvested data (e.g., a tweet) may not provide enough information and meaning to facilitate automatic classification [17]. Second, the collected data may not be representative of the entire population—this challenge is especially pertinent as social media users are often younger, more educated, and urban-dwelling with higher incomes [17]. Furthermore, while coding for geographic origin may resolve certain limitations, not all profile accounts on networking sites contain geographic information; even so, visible geographic information cannot be verified for accuracy.

It is thus worthwhile to explore data mining sources that track IP addresses, or techniques to monitor social media activity on mobile phones [17]. Alternatively using GPS monitoring – using GIS systems that are either embedded in smart phones or attached to independent devices can provide supplemental information. One question that has not been resolved is whether participants need to provide explicit consent or whether the fact they are using publically available communication tools renders their information available and subsequently usable.

4.2 Study Limitations

Our search strategy was not as streamlined as we had planned. Not having the ability to choose a MeSH term as distinct as "social media", "infoveillance", or "Twitter", meant that time was wasted sifting through irrelevant publications. While we extended our search to include conference proceedings as well, the publications retrieved were not of value. Rather, the references in retrieved publications provided direction to relevant crucial proceedings. However, some of the publications we retrieved through hand searching were not retrieved through direct bibliographic database searching.

4.3 Research Directions

In this section, we outline key recommendations that we believe are essential to forging new opportunities in data mining and collaborative analysis within user-driven content sharing paradigms. This will enable the full realization of the significant potential of patient engagement and information sharing and may help transform healthcare as a whole.

Target Audiences. The initial task for any text mining solutions is identifying entities of interest from the relevant textual content. This is often achieved through the use of natural language processing techniques. Each social media has a specific target audience with unique engagement behavior specific to that platform. Information harvested through social networking sites may not be representative of the population

at large. It is therefore important that data mining solutions take into account demographic characteristics of audiences within individual platforms.

User Interactions: Every social networking platform has a set of rules governing how its users interact with one another. For example, some social media platforms such as Twitter and Facebook enable real-time interactions between users while YouTube tends to be less interactive. As such, the types of interactions will determine when and how often data must be collected in order to derive any meaningful information.

Information Access: Social media platforms are already compiling fine-grained user-generated content based on individuals' online activities. While the means for deciphering what is relevant through information mining already exists and have proven extremely successful considering the amount of money companies are willing to pay to have such kind of access, such personal and information-rich content is not often publicly available. More open-source social networking solutions are therefore needed to facilitate any meaningful data mining solutions beyond the basic alert systems discussed in this paper.

Evidence-Based Medicine: Research is needed to identify effective ways of embedding evidence within social media platforms that could support monitoring positive impact on desired behavior changes and allowing users to share/compare experiences and provide support. Given the broad range of users, there is also a need to provide levels of detail regarding the evidence itself so that meaning information could be mined.

Knowledge Representation: Finally, mining social media content for medical information can only succeed if we recognize the role of ontologies in knowledge management and knowledge discovery. Ontology offers significant benefit to knowledge harvesting in social networking platforms as it facilitates data pruning and can help accelerate the discovery of meaningful information.

Summary. Evolving the use of infodemiology in healthcare will involve the examination of: (1) user interactions, which may determine the time and frequency of data collection; (2) information access, which may require the creation of more open-source social networking sites to facilitate more meaningful data mining solutions; (3) evidence-based medicine, which will allow users to share and compare experiences and provide support from within a social media platform; and (4) knowledge representation, which can allow the mining of social media for medical information, knowledge management and knowledge discovery. Health and healthcare might soon be achieved at the click of a mouse.

References

1. Evans, D.: Social Media Marketing: An Hour a Day. Wiley Publishing, Indiana (2008)
2. McNab, C.: What Social Media Offers to Health Professionals and Citizens. Bulletin of the WHO 87, 566 (2009)
3. Bacigalupe, G.: Is There a Role for Social Technologies in Collaborative Healthcare? Families, Systems & Health 29, 1–14 (2011)

4. Culotta, A.: Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: 1st Workshop on Social Media Analytics (SOMA 2010), Washington DC (2010)
5. Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q.S., Baker, M.: Interpreting Google Flu Trends Data for Pandemic H1N1 Influenza: The New Zealand Experience. Euro. Surveill. 14, 19386 (2009)
6. Chew, C., Eysenbach, G.: Pandemics in the Age of Twitter: Content Analysis of Tweets During the 2009 H1N1 Outbreak. PLoS ONE 5 (2010)
7. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Annals of Internal Medicine 151, 264–269 (2009)
8. Achrekar, H., Avinash, G., Lazarus, R., Yu, S., Liu, B.: Predicting Flu Trends using Twitter Data. In: International Workshop on CPNS, Shanghai, China (2011)
9. Chen, L., Achrekar, H., Liu, B., Lazarus, R.: Vision: Towards Real Time Epidemic Vigilance through Online Social Networks: Introducing SNEFT. In: 1st ACM Workshop on Mobile Cloud Computing, San Francisco, California (2010)
10. Corley, C.D., Mikler, A.R., Singh, K.P., Cook, D.J.: Monitoring Influenza Trends through Mining Social Media. In: International Conference on Bioinformatics and Computational Biology, Las Vegas, Nevada (2009)
11. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Text and Structural Data Mining of Influenza Mentions in Web and Social Media. International Journal of Environmental Research and Public Health 7, 596–615 (2010)
12. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Using Web and Social Media for Influenza Surveillance. Advances in Experimental Medicine and Biology 680, 559–564 (2010)
13. de Quincey, E., Kostkova, P.: Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter. In: Kostkova, P. (ed.) eHealth 2009. LNICST, vol. 27, pp. 21–24. Springer, Heidelberg (2010)
14. Eysenbach, G.: Infodemiology and Infoveillance Tracking Online Health Information and Cyberbehavior for Public Health. American Journal of Preventive Medicine 40, S154–S158 (2011)
15. Lampos, V., Cristianini, N.: Tracking the Flu Pandemic by Monitoring the Social Web. In: Int. Workshop on CIP, Elba Island, Italy (2010)
16. Lopes, L.F., Zamite, J.M., Tavares, B.C., Couto, F.M., Silva, F., Silva, M.J.: Automated Social Network Epidemic Data Collector. In: INForum (2009)
17. Signorini, A., Segre, A.M., Polgreen, P.M.: The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PloS One 6, e19467 (2011)
18. Wilson, K., Brownstein, J.S.: Early Detection of Disease Outbreaks Using the Internet. CMAJ: Canadian Medical Association Journal 180, 829–831 (2009)
19. Brownstein, J.S., Freifeld, C.C., Madoff, L.C.: Digital Disease Detection–Harnessing the Web for Public Health Surveillance. The New England Journal of Medicine 360, 2153–2155, 2157 (2009)