COACTION PUBLISHING

A review of automated text classification in event-based biosurveillance

Manabu Torii*, Jeffrey Collmann, David Hartley and Noele Nelson

Georgetown University Medical Center, Washington, DC, USA

Objective
The objective of this literature review is to identify current challenges in document classification for event-based biosurveillance and consider the necessary efforts.

Introduction
Event-based biosurveillance monitors diverse information sources for the detection of events pertaining to human, plant and animal health using online documents, such as news articles, newsletters and blogs (1). Machine learning techniques have been successfully used for automated document classification, an important step in filtering source information (2–15).

Methods
We review studies on document classification using machine learning for event-based biosurveillance and comparatively summarize them for close examination.

Results
Table 1 lists relevant studies we identified. These studies differ in target regions, languages, event types and surveillance criteria, as well as classification methods. This diversity illustrates the complementarity of all the approaches.

Conclusions
Common challenges shared by these methods include detection of rare events and practical evaluation of the employed methods. The comparative advantages of each method remain unclear because of the lack of benchmark data. A community effort is necessary to develop an event ontology and benchmark corpora.

Table 1. Selected studies

Program	*Reference*	*Primary source*	*Methods*
Argus I	Lehner et al. (10)	News	NB
Argus II	Torii et al. (12)	News	NB, SVM
Biocaster	Conway et al. (4)	News	DT, NB, SVM
	Doan et al. (6)	News	NB, SVM
EpiSpider	Tolentino et al. (11)	ProMed	NB
GPHIN	Blench (3)	News	'proprietary'
Healthmap	Freifeld et al. (8)	News	NB
InSTEDD	http://instedd.org/evolve	News	NB, SVM
-	Aramaki et al. (2)	Twitter	SVM
-	Culotta (5)	Twitter	LR
-	von Etter et al. (7)	News	NB, SVM
-	Lampos et al. (9)	Twitter	Bolasso
-	Signorini et al. (15)	Twitter	SVM
-	Zhang et al. (14)	News	kNN, NB, SVM
-	Zhang and Liu (13)	ProMed	NB, SVM

Abbreviations: NB, naive Bayes.

Keywords
biosurveillance; text classification; machine learning

References

1. Hartley DM, Nelson NP, Walters R, Arthur R, Yangarber R, Madoff L, et al. The landscape of international event-based biosurveillance. Emerg Health Threats J. 2010;3:e3.
2. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using twitter. EMNLP. 2011:1 568–76.
3. Blench M. Global Public Health Intelligence Network (GPHIN). AMTA. 2008:299–303.
4. Conway M, Doan S, Kawazoe A, Collier N. Classifying disease outbreak reports using n-grams and semantic features. SMBM. 2008:29–36.
5. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. SOMA; 2010.
6. Doan S, Kawazoe A, Conway M, Collier N. Towards role-based filtering of disease outbreak reports. J Biomed Inform. 2009;42:773–80.
7. von Etter P, Huttunen S, Vihavainen A, Vuorinen M, Yangarbe R. Assessment of utility in web mining for the domain of public health. LOUHI. 2010:29–37.
8. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc. 2008;15:150–7.
9. Lampos V, Bie TD, Cristianini N. Flu detector—tracking epidemics on Twitter. ECML PKDD. 2010:599–602.
10. Lehner P, Worrell C, Vu C, Mittel J, Snyder S, Schulte E et al. An application of document filtering in an operational system. Inform Process Manage. 2010:46.
11. Tolentino H, Kamadjeu R et al. Scanning the emerging infectious diseases horizon—visualizing ProMED emails using EpiSPIDER. Adv Dis Surveill. 2007; 2:1.
12. Torii M, Yin L, Nguyen T, Mazumdar CT, Liu H, Hartley DM, et al. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. Int J Med Inform. 2011;80:56–66.
13. Zhang Y, Liu B. Semantic text classification of emergent disease reports. PKDD. 2007:629–37.
14. Zhang YL, Dang Y, Chen H, Thurmond M, Larson C. Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems. 2009;47:508–17.
15. Signorini A, Segre AM, Polgreen PM. 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. 2011;6:e19467.

***Manabu Torii**
E-mail: manabu.torii@gmail.com

Emerging Health Threats Journal 2011. © 2011 M. Torii et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Emerging Health Threats Journal 2011, **4**: 11057 - DOI: 10.3402/ehtj.v4i0.11057
(page number not for citation purpose)