ISDS 2012 Conference Abstracts

Content Analysis of Syndromic Twitter Data

Bethany Keffala*[1], Mike Conway[2], Son Doan[2] and Nigel Collier[3]

[1]Linguistics, University of California, San Diego, La Jolla, CA, USA; [2]University of California, San Diego - Division of Biomedical Informatics, La Jolla, CA, USA; [3]National Institute of Informatics, Tokyo, Japan

Objective

We present an annotation scheme developed to analyze syndromic Twitter data, and the results of its application to a set of respiratory syndrome-related tweets [1]. The scheme was designed to differentiate true positive tweets (where an individual is experiencing respiratory symptoms) from false positive tweets (where an individual is not experiencing respiratory symptoms), and to quantify more fine-grained information within the data.

Introduction

The popularity of Twitter, a social-networking service, creates the opportunity for researchers to collect large amounts of free, localizable data in real-time. Data takes the form of short, user-written messages, and has been employed for general syndromic surveillance [2] and surveillance of public attitudes toward the H1N1 flu outbreak [3]. Accessibility of tweets in real-time makes them particularly appropriate for use in early warning systems. Data collected through keyword search contains a significant amount of noise, however, annotation can help boost the signal for true positive tweets.

Methods

The annotation scheme was developed based on information relevant for early warning systems (e.g. who is experiencing symptoms, and when) as well as other information present in the tweets (e.g. aspirations regarding symptoms, or abuse of substances such as cough syrup). Categories included Experiencer: Self/Other, Temporality: Current/Non-Current, Sentiment: Positive/Negative, Information: Providing/Seeking, Language: Non-English, Aspiration, Hyperbole, and Substance Abuse. All categories with the exception of Language and Substance Abuse were defined in reference to diseases or symptoms. The scheme was applied to 1,100 respiratory syndrome-related tweets (544 false positive, 556 true positive) from a previously collected corpus of syndromic twitter data [2]. Inter-annotator agreement was calculated for 9% of the data (100 tweets).

Results

Inter-annotator agreement was generally good, however certain categories had lower scores. Categories for Experiencer, Temporality, Sentiment: Negative, Information: Providing, and Language all had Kappa values above .9, Sentiment: Positive, Aspiration, and Substance abuse had Kappa values above .7, and Information: Seeking and Hyperbole had Kappas above .6. There was good separation between true positive tweets and false positive tweets, especially for the Experiencer: Self, Temporality: Current, Sentiment: Negative, Aspiration, Hyperbole, and Substance Abuse categories (see Table). True positive data were more likely to belong to any category except Information: Providing, and Substance Abuse, in which cases false positive tweets had greater likelihood of category inclusion. Within the true positive data, we found that users were more likely to reference symptoms that they themselves were currently experiencing than they were to reference another person's symptoms or non-current symptoms. Sentiment was largely negative, and there was significant use of aspiration and hyperbole.

Conclusions

Future work will apply the scheme to other syndromes, including constitutional, gastrointestinal, neurological, rash, and hemorrhagic.

	% True Positive Tweets	% False Positive Tweets
Experiencer: Self	98.2	0.4
Temporality: Current	98.7	0.2
Sentiment: Negative	79.7	1.7
Information: Providing	0.7	2.8
Language: Non-English	2.7	1.3
Aspiration	11.0	0.2
Hyperbole	18.3	0.2
Substance Abuse	1.3	8.1

Table 1. Percentages of tweets included in each category.

Keywords

social media; surveillance; respiratory syndrome

References

1. N. Collier, R. Matsuda Goodwin, J. McCrae, S. Doan, A. Kawazoe, M. Conway, A. Kawtrakul, K. Takeuchi, D. Dien. (2010). "An ontology-driven system for detecting global health events", Proc. 23rd International Conference on Computational Linguistics (COLING), Beijing, China, August 23-27, pp. 215-222, available from http://aclweb.org/anthology/C/C10/C10-1025.pdf.
2. Collier, N. & Doan, S. (2011). "Syndromic Classification of Twitter Messages", Proc. eHealth 2011, Malaga, Spain. November 21-23.
3. Chew, C. & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE 5(11): e14118.

*Bethany Keffala
E-mail: bkeffala@ucsd.edu