Asthma Surveillance Using Social Media Data

Wenli Zhang[1], Sudha Ram[1], Mark Burkart[2], Max Williams[2], and Yolande Pengetnze[2]
University of Arizona[1], PCCI-Parkland Center for Clinical Innovation[2]
{wenlizhang, sram}@email.arizona.edu, {MARK.BURKART, MAX.WILLIAMS, YOLANDE.PENGETNZE}@phhs.org

Introduction and background

Asthma is a chronic disease that affects people of all ages. In the United States (US), more than 25 million people are known to have asthma [1]. Asthma can cause recurring periods of wheezing, chest tightness and shortness of breath, and has no cure. Annually it results in more than 2 million emergency department visits, half a million hospitalizations and 3,500 deaths in the US. Accurate and timely asthma surveillance data could guide intervention efforts to improve the quality of asthma patient's life and reduce the societal burden of asthma. According to the Centers for Disease Control and Prevention (CDC), current asthma surveillance is done by collecting data at the national and the state level [2]. Limit municipal level data are available from CDC. Notoriously, such data have a lag-time of months or years. The most recent national and state asthma prevalence data was collected in 2013 [3]. Thus, current surveillance systems are neither suitable for timely asthma interventions, nor enough for health organizations to proactively prepare for asthma emergency department visits in real time.

Social media platforms for microblogging, e.g., Twitter, are widely used today by a large number of people. Disease self-reporting, e.g., "I had an asthma attack!" are not uncommon on social media. There has been increasing interest in using social media data for surveillance and predictive analytics in health care because such data can provide nearly instant access to real-time latent population characteristics. Culotta [4] proposed a method to identify influenza-related tweets to correlate them with CDC statistics. Study from Collier [5] show a high degree of correlation between pre-diagnostic social media signals and diagnostic influenza case data. Broniatowski [6] developed a pipeline classification system to identify relevant data for influenza surveillance. However, increasing evidence suggests that many of the predictions and analyses produced misrepresent the real world [7]. The main challenges arise from several different types of noise in the datasets, including: (a) use of loosely structured informal language: e.g.,

misspellings and emoticons tend to bias machine learning techniques toward misclassification of text; (b) anomalous media spikes: e.g., people may include flu related terms in their tweets, however some of these maybe from users who simply retweet flu news stories – these tweets do not necessarily reflect actual disease affliction; and (c) use of misleading terms and phrases: e.g., tweets indicating awareness of flu, e.g., "Hope I won't get flu" – these messages are clearly about the flu but not about an infection.

Our research objective is to develop an effective methodology to extract signal from social media data and use it to provide accurate and timely asthma surveillance information at states and municipal levels. This work is different from extant studies that typically predict the spread of contagious diseases using social media. The results of this study have the potential to enable health organizations and public health entities to respond to chronic conditions, like asthma, in real time. This in turn implies that health organizations can appropriately plan for staffing and equipment in a flexible manner.

Methods

We introduce a novel methodology of collecting and using social media data, i.e., Twitter data, for asthma surveillance at the state and municipal level in near real-time. A combination of different Natural language processing (NLP) feature generation methods along with machine learning classifiers are proposed to identify and remove various types of noise from social media text. Geo-location filter are used to map tweets in different regions. The US adult asthma prevalence data from CDC and real-world asthma hospital data from the Dallas-Fort Worth area were used to evaluate the results.

	# of tweets	Collection period	Geographical area	# of keywords	Keywords examples
Asthma Twitter Dataset	5,513,368	11/1/2013 – 6/30/2014	Global	18	Asthma, inhaler

Table 1. Asthma-related Twitter dataset

Twitter provides APIs to access large volumes of user-generated text as well as automatically generated information (e.g., content creation time, users' geo-location). We collected asthma-related Twitter stream containing one or more of 18 related keywords that were suggested by the clinical collaborators from Parkland Center for Clinical Innovation (PCCI) [8]. A large dataset of more than 5 million asthma-related tweets was collected over a period of approximately 6 months (Table 1).

To identify various types of noise, our signal extraction methodology included several processing steps: text preprocessing, feature extraction, and classification (Figure 1). The goal is twofold: first, to isolate various types of linguistic noise; and second, to distinguish asthma-relevant (i.e., actual asthma infections/symptoms) from asthma-irrelevant (i.e., media spikes and misleading terms/phrases) text.

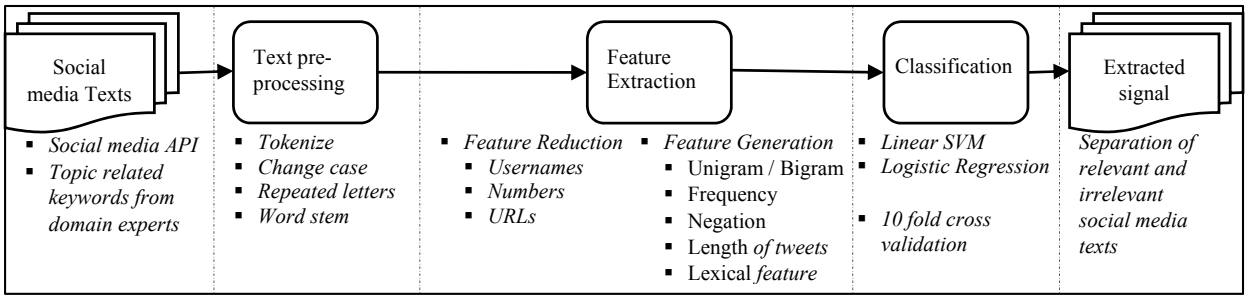

Figure 1. Methodology for signal extraction from social media text

Social media users tend to use casual language. Text pre-processing can effectively reduce lexical noise. Text pre-processing includes: (a) tokenization; (b) case-folding; (c) removing repeated letters and (d) word stemming. Feature extraction is a process used to transform social media text into numerical vectors for classification using machine learning techniques which includes feature reduction and generation. Feature reduction is important because it can improve the efficiency of training a classification model by decreasing the size of the effective vocabulary and classification accuracy can be increased by eliminating noisy features. Usernames (e.g., @alexwu); numbers, currency and percentages; URLs were replaced by a single token (_USERNAME), (_NUMBER) or (_URL) respectively. Feature generation is aimed at creating and selecting a subset of word tokens as features for machine learning techniques. We explore the use of unigrams, bigrams, negations, lexical features (part-of-speech tags), length of text, and their combinations as features. A training dataset, containing 4,500 tweets, was created from the asthma Twitter dataset and each tweet was manually labeled by three researchers as "asthma relevant" or "asthma irrelevant" [8]. Classification methods were used to separate asthma-relevant and irrelevant tweets and 10 fold cross validation were executed to evaluate the performance (Table 2) [9].

	Accuracy	Asthma relevant			Asthma irrelevant		
		precision	recall	f1	precision	recall	f1
Linear SVM	0.872	0.605	0.630	0.617	0.927	0.920	0.923
Logistic Regression	0.886	0.667	0.603	0.633	0.924	0.941	0.932

Asthma training data set; 10 fold cross validation

Table 2. Performance of signal extraction from social media text

Geo-location extraction is an important step for studies that involve geographic analyses. The geographic location of each tweet is identified via two fields: coordinates and location. Coordinates indicate the longitude and latitude of the tweet's location, e.g., {"coordinates": [-97.510, 35.465]}. Locations indicate the cities and states where the users, who posted tweets, reside, e.g., {"location": "San Francisco, CA"}. This information is collected from Twitter users' profiles. However, there is considerable ambiguity in the location information, e.g., "I am somewhere on the earth". Geo-location extraction was performed by using (1) latitude and longitude mapping for the coordinates field and (2) regular expressions analysis for the location field, i.e., matching state names or postal abbreviations on the US states (e.g., Texas or TX), followed by matching city names (e.g. Dallas). Users' self-reported time zone information were used to distinguish different cities with same city names (e.g., Arlington: city in Texas, Central Time Zone and Arlington: city in Washington DC, Eastern Time Zone).

Results and Discussion

At the state level, we report the relationship between the aggregate dataset from CDC on US adult asthma prevalence (number with current asthma in 2013) [3] and our Twitter asthma dataset. Among all the global asthma-related tweets, 3.10% (171,165 / 5,513,368) of the tweets contained geographic coordinates. There were 91.03% (5,019,319 / 5,513,368) tweets containing location information in the user profile. We were able to identify location for approximately 18.85% (63,093 /517,342) tweets as one of 50 US state names. Based on the results (Table 3) asthma related tweets in different US states correlates strongly with CDC asthma prevalence data. The correlation to the actual asthma prevalence significantly improved from 0.303 to 0.692 after using our signal extraction methods. The results demonstrate that our method is very effective for surveillance of asthma prevalence at a State level.

		After signal extraction	Before signal extraction
Asthma Prevalence 2013	**Pearson Correlation**	0.692**	0.303*
	sig.	0	0.029
	N	50	50

**. Correlation is significant at the 0.01 level *. Correlation is significant at the 0.05 level

Table 3. Correlation between asthma prevalence and Twitter asthma data set

At the municipal level, we report on the correlation between the de-identified aggregate data on hospital visits for asthma as a primary diagnosis (International Classification of Disease Ninth [ICD9] code 493.00 to 493.99) from 28 hospitals/health organizations in the Dallas–Fort Worth area (between November 2013 and Jun 2014) and our Twitter asthma dataset. 18,583 tweets of these were either mapped to the Dallas-Fort Worth area with coordinates or the location fields were recognized as one of 70 major cities (population > 10,000) in the Dallas-Fort Worth area. Results (Table 4) show that the same methodology (described above) accurately captures asthma incidence at the level of a municipality. After signal extraction and geo-location filtering, Twitter asthma dataset was highly correlated with the asthma hospital visits in the Dallas-Fort Worth area both at monthly and daily levels.

		after signal extraction	before signal extraction
Asthma hospital visits Dallas-Fort Worth Aggregate by month-year	**Pearson Correlation**	0.746*	0.449
	sig.	0.021	0.225
	N	8	8

Table 4(a). Correlation between the monthly number of asthma tweets and the same month number of asthma ED/hospital visits in the Dallas-fort worth area

		after signal extraction	before signal extraction
Asthma hospital visits Dallas-Fort Worth Aggregate by day	**Pearson Correlation**	0.281**	0.340**
	sig.	<0.001	<0.001
	N	244	244

**. Correlation is significant at the 0.01 level *. Correlation is significant at the 0.05 level

Table 4(b). Correlation between the daily number of asthma tweets and the same day number of asthma ED/Hospital visits in the Dallas-Fort Worth area

Implications and Conclusions

In this study, we developed a comprehensive methodology to use social media data and extract signal from social media text for the purpose of accurate and timely chronic disease surveillance, specifically asthma. The results can be helpful for public health surveillance, emergency department preparedness, and targeted patient interventions. For the future work: population biases vary across different social media platforms, i.e., some social media sites are more appealing to teenagers. These biases should be fully acknowledged and corrected. More sophisticated methods such as semi-supervised could be developed to further improve the efficiency of our signal extraction methods. We are also in the process of developing a prediction model for accurately determining the number of people likely to show up in the ED and/or clinics for asthma related complications using both social media and streaming air quality sensor datasets.

Reference

[1] Centers for Disease Control and Prevention. 2014, "About the Morbidity and Mortality Weekly Report (MMWR) Series." [Online] Available: http://www.cdc.gov/mmwr/about.html. Accessed at: Sep 15, 2014.

[2] Centers for Disease Control and Prevention. "Asthma Data, Statistics, and Surveillance." [Online] Available: http://www.cdc.gov/ asthma /asthmadata.htm. Accessed at: Aug 9, 2014.

[3] Centers for Disease Control and Prevention. "Most Recent Asthma Data." [Online] Available: http://www.cdc.gov/asthma/most_recent_data_states.htm. Accessed at: Aug 9, 2014.

[4] Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." In Proceedings of the first workshop on social media analytics, pp. 115-122. ACM, 2010.

[5] Collier, Nigel, Nguyen Truong Son, and Ngoc Mai Nguyen. "OMG U got flu? Analysis of shared health messages for bio-surveillance." J. of Biomedical Semantics 2, no. S-5, S9, 2011.

[6] Broniatowski, David A., Michael J. Paul, and Mark Dredze. "National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic." PloS one vol. 8, no.12, e83672, 2013.

[7] Ruths, D., & Pfeffer, J. 2014. “Social media for large studies of behavior”. Science, 346(6213), 1063-1064.

[8] Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. 2015. “Predicting Asthma-Related Emergency Department Visits Using Big Data”. IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, 2015(7).

[9] Zhang, W., Ram, S. 2015. "A Comprehensive Methodology for Extracting Signal from Social Media Text Using Natural Language Processing and Machine Learning". 25th Workshop on Information Technologies and Systems (WITS). [Under review]