ISDS 2012 Conference Abstracts

#wheezing: A Content Analysis of Asthma-Related Tweets

Gwendolyn Gillingham*[1], Michael A. Conway[2], Wendy W. Chapman[2], Michael B. Casale[3] and Kathryn B. Pettigrew[3]

[1]Linguistics, UCSD, La Jolla, CA, USA; [2]UCSD - Division of Biomedical Informatics, La Jolla, CA, USA; [3]West Health Institute, La Jolla, CA, USA

Objective

We present a Content Analysis project using Natural Language Processing to aid in Twitter-based syndromic surveillance of Asthma.

Introduction

Recently, a growing number of studies have made use of Twitter to track the spread of infectious disease. These investigations show that there are reliable spikes in traffic related to keywords associated with the spread of infectious diseases like Influenza [1], as well as other Syndromes [2]. However, little research has been done using Social Media to monitor chronic conditions like Asthma, which do not spread from sufferer to sufferer. We therefore test the feasibility of using Twitter for Asthma surveillance, using techniques from NLP and machine learning to achieve a deeper understanding of what users Tweet about Asthma, rather than relying only on keyword search.

Methods

We retrieved a large volume of Tweets from the Twitter API. Search terms included "asthma," and several misspellings of that word; terms for common medical devices associated with Asthma such as "inhaler" and "nebulizer"; and names of prescription drugs used to treat the condition, including "albuterol" and "Singulair." A randomly sampled subset of these Tweets (N=3511) was annotated for content, based on an annotation scheme that coded for the following elements: the Experiencer of Asthma symptoms (Self, Family, Friend, Named Other, Unidentified, and All-Non-Self, which was the union of these last four categories); aspects of the type of information being conveyed by each Tweet (Medication, Triggers, Physical Activity, Contacting of a Medical Practitioner, Allergies, Questions, Suggestions, Information, News, Spam); as well as Negative Sentiment, Future temporality, and Non-English content. Further details on the annotation scheme used can be found at http://idiom.ucsd.edu/~ggilling/annotation.pdf. Inter-annotator agreement on a subset of the Tweets (N=403) fell in an acceptable range for all categories (Cohen's Kappa >0.6). Once annotation was complete, the Tweets' texts were stemmed and converted into vectors of unigram and bigram counts. These were then stripped of sparse terms (all those words appearing in fewer than 1 in 200 Tweets), which left multi-dimensional vectors consisting of the counts of the remaining words in all Tweets. Statistical machine-learning classifiers including K-nearest neighbors, Naive Bayes and Support Vector Machines were then trained on the unigram and bigram models.

Results

SVM with 10-fold cross-validation achieved greatest prediction accuracy with the unigram model, as shown in Table 1. Categories that showed the greatest reduction in classification error using the unigram model were Non-English, Self, All-Non-Self, Medication, Symptoms and Spam. The majority of these categories showed very high Precision, as well as fairly high Recall for the unigram model. Unexpectedly, the bigram model faired far worse than the Unigram model, which suggests that individual words in these Tweets were more reliably predictive of content than pairs of words, which occurred less frequently.

Conclusions

Text-classification increases the utility of Twitter as a data-source for studying chronic conditions such as Asthma. Using these methods, we can automatically reject Tweets that are non-English or Spam. We can also determine who is experiencing symptoms: the Twitter user or another individual. Fairly simple models are able to predict with good certainty whether a user is talking about their Symptoms, their Medication, or Triggers for their Asthma, as well as whether they are expressing Negative sentiment about their condition. We demonstrate that Social Media such as Twitter is a promising means by which to conduct surveillance for chronic conditions such as Asthma.

Table 1: Performance of Classifiers on Unigram and Bigram Models

Dimension	Baseline Error (= 1-Majority Classification/N)	Unigram Model Error	Bigram Model Error	Unigram Precision	Unigram Recall
Non-English	0.19	0.07	0.17	0.9	0.82
Self	0.22	0.16	0.18	0.94	0.76
All-Non-Self	0.19	0.14	0.17	0.94	0.59
Medication	0.15	0.098	0.16	0.89	0.77
Symptoms	0.21	0.16	0.17	0.9	0.76
Spam	0.07	0.055	0.17	0.93	0.43

Keywords

social media; natural language processing; asthma; content analysis

Acknowledgments

This work was financially supported by the West Wireless Health Institute and iDASH Summer Internship program (NIH U54HL108460).

References

1. Chew, C. & Eysenbach, G. 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets in the H1N1 Outbreak. PLoS ONE 5(11): e14118.
2. Collier, N. & Doan, S. 2011. Syndromic Classification of Twitter Messages. Proc. eHealth 2011, Malaga, Spain. November 21-23.

*Gwendolyn Gillingham
E-mail: gwen.gillingham@ling.ucsd.edu