OJPHI

ISDS 2014 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

User-Customizable Health Pattern Detector Framework: Twitter Analysis Example

Lianna M. Hall*, Kevin K. Nam, Jason Thornton, Marianne DeAngelus and Timothy J. Dasey

Informatics and Decision Support, MIT Lincoln Laboratory, Lexington, MA, USA

Objective

To demonstrate a framework for user-customizable text processing that can improve the efficiency and effectiveness of mining text for biosurveillance, with initial application to Twitter.

Introduction

Early detection of a disease outbreak using pre-diagnostic textual data is available in biosurveillance systems with the integration of data such as chief complaints. Social media has been identified as an additional pre-diagnostic data source of interest[1]. Textual data analysis in public health is usually based on a keyword search and often involves a complex Boolean combination of terms that produce results with many false alarms. Epidemiologists may wish to query the data differently based on the event of interest, yet the process is laborious to weed out uninteresting content. Specialized detectors that decide on the topical relevance of keyword search usually require developers to adapt methods to new uses, which is a time- and cost-prohibitive activity. Users need the ability to rapidly build text content detectors on their own.

Methods

A generalizable detector framework called Customizable Pattern Analytics (CPA) was adapted and tested with the Twitter biosurveillance data mining application. CPA was originally developed for detecting features in videos[2], but has a general purpose mathematical framework that allows migration to other data discrimination problems. CPA automatically reconfigures multiple stages of a detector processing chain (e.g. feature selection and classification) based on binary feedback from the user on the utility of returned results. It does so by computing a wide range of features about the data, and adjusting the feature weighting and the decision boundary on the combined features based on user feedback. The result is a user-built detector that can be specific to a situation.

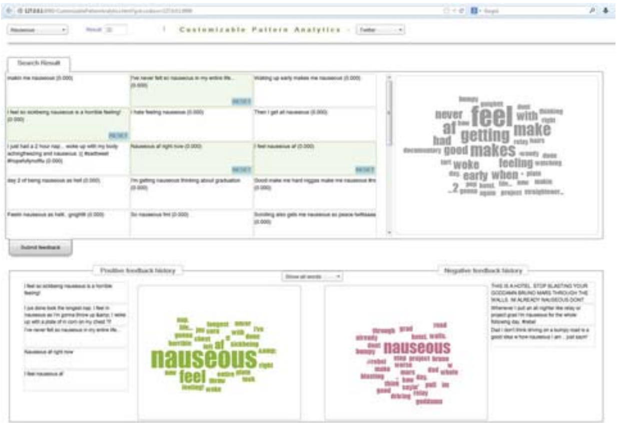
For Twitter processing, CPA analyzes many characteristics of each Tweet that is returned from a keyword search, including term frequencies, common word combinations, content flags, and metadata (e.g. location). A user interface transparently shows ranked examples of the returned Tweets of suspected relevance to the user. The user can select examples as either relevant or irrelevant, and the interface progressively displays a new set of options based on an underlying reengineering of the detector by CPA.

Results

The figure shows an example user interface screenshot. The application is available for demonstration. Performance curves (i.e. true vs. false positive rate) show that CPA achieves superior performance than that of keyword search alone or from one specific type of text analysis. Importantly, the type of text processing to apply is varied based on the particular keywords used in the search. CPA can select the most productive combination of text processing methods to apply that best match the user-supplied yes/no labels. Empirical trials have demonstrated that there is a significant accuracy boost with as few as 5 yes/no labels, and that this accuracy approaches its upper limit after several dozen feedback labels.

Conclusions

Application of CPA to Twitter data analysis was demonstrated with superior performance to simpler text analysis and with the versatility to be applied to a wide range of microblog processing tasks. These methods should be directly applicable to similar text processing tasks (e.g., chief complaints). Automatic or user-guided keyword expansion methods are being investigated to extend the capability. Extensions to other text processing problems are imagined, such as for electronic health record analysis.



Waking up early makes me nauseous (0.000)

I hate feeling nauseous (0.000)

Then I get all nauseous (0.000)

Nauseous af right now (0.000)

I feel nauseous af (0.000)

I'm getting nauseous thinking about graduation (0.000)

So nauseous fml (0.000)

THIS IS A HOTEL. STOP BLASTING YOUR GODDAMN BRUNO MARS THROUGH THE WALLS. IM ALREADY NAUSEOUS DONT

Example user interface

Keywords

visual analytics; user adaptable search; social media

Acknowledgments

This work is sponsored by the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

References

1. CDC. Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014. MMWR 2014; 63(32); 681-685.
2. Thornton J, DeAngelus M, Chan M. Online Customization of Video Detection Capabilities. IEEE Int'l Conf. on Security Technology; 2014 Oct 13-16; Rome, Italy. Accepted for presentation.

*Lianna M. Hall
E-mail: lianna.hall@ll.mit.edu