Use of IDATA Capabilities for Social Media Analytics

Thi Uyen Tran and Daniel Nakada

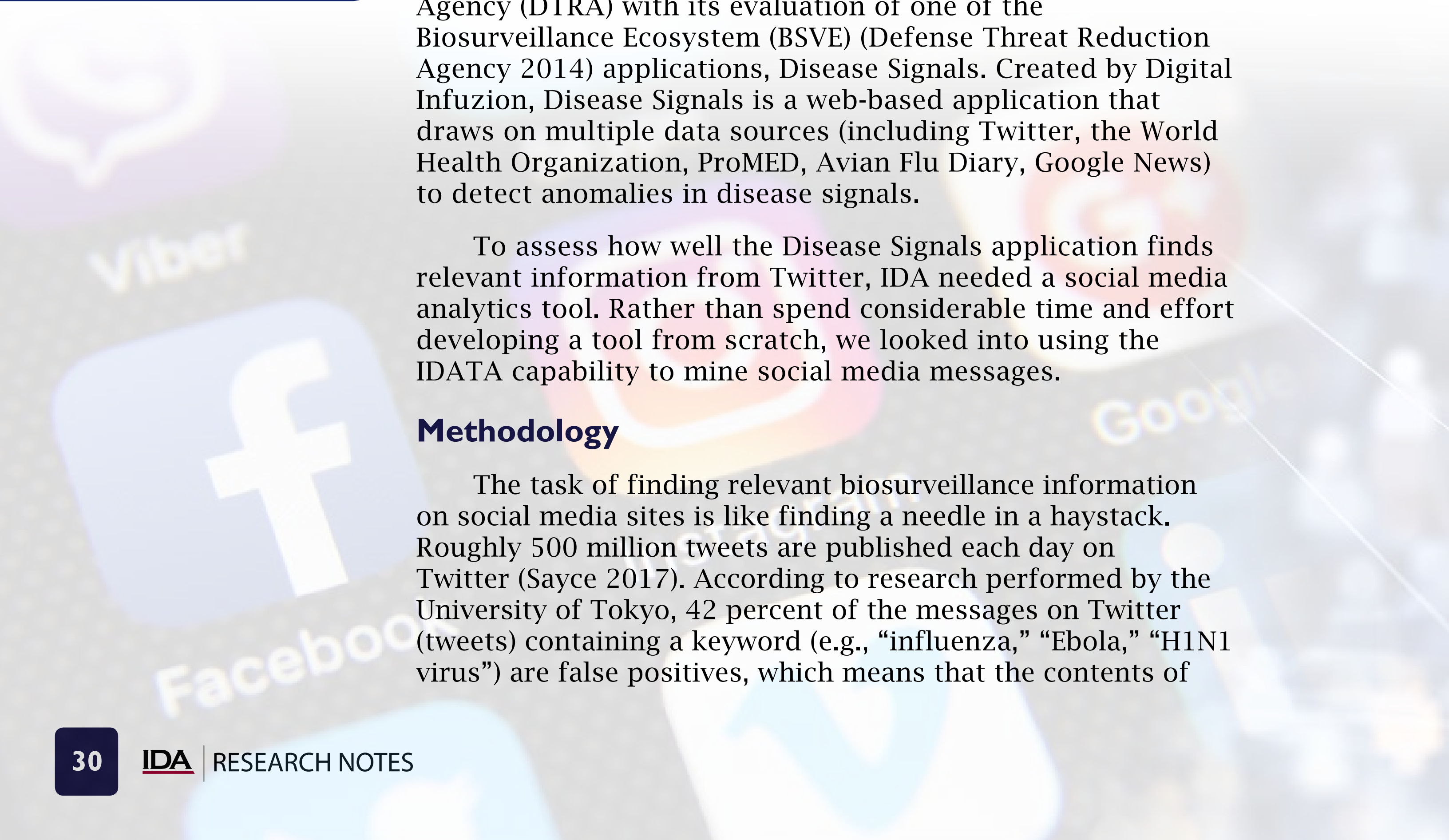
The Challenge: The Defense Threat Reduction Agency needed to understand how well a biosurveillance application finds relevant disease information on social media.

Social media sites produce information that research communities can use to improve responses to national security and public health problems, such as measuring public anxiety after a natural disaster.

Background

In recent years, the emergence of social media sites such as Facebook, Twitter, Snapchat, and LinkedIn has fundamentally shifted the way people communicate and share information. Today, political views, religious beliefs, and even personal health status can be transmitted easily at near-real-time speed. This phenomenon produces a wealth of information that research communities can use to improve responses to national security and public health problems, such as measuring public anxiety after a natural disaster (Doan, Ho Vo, and Collier 2011), detecting an earthquake through the "social sensor" (Sakaki, Okazaki, and Matsuo 2010), monitoring bribery or violence during an election (Draxler 2014), or detecting and tracking infectious disease outbreaks.

IDA was asked to assist the Defense Threat Reduction Agency (DTRA) with its evaluation of one of the Biosurveillance Ecosystem (BSVE) (Defense Threat Reduction Agency 2014) applications, Disease Signals. Created by Digital Infuzion, Disease Signals is a web-based application that draws on multiple data sources (including Twitter, the World Health Organization, ProMED, Avian Flu Diary, Google News) to detect anomalies in disease signals.



To assess how well the Disease Signals application finds relevant information from Twitter, IDA needed a social media analytics tool. Rather than spend considerable time and effort developing a tool from scratch, we looked into using the IDATA capability to mine social media messages.

Methodology

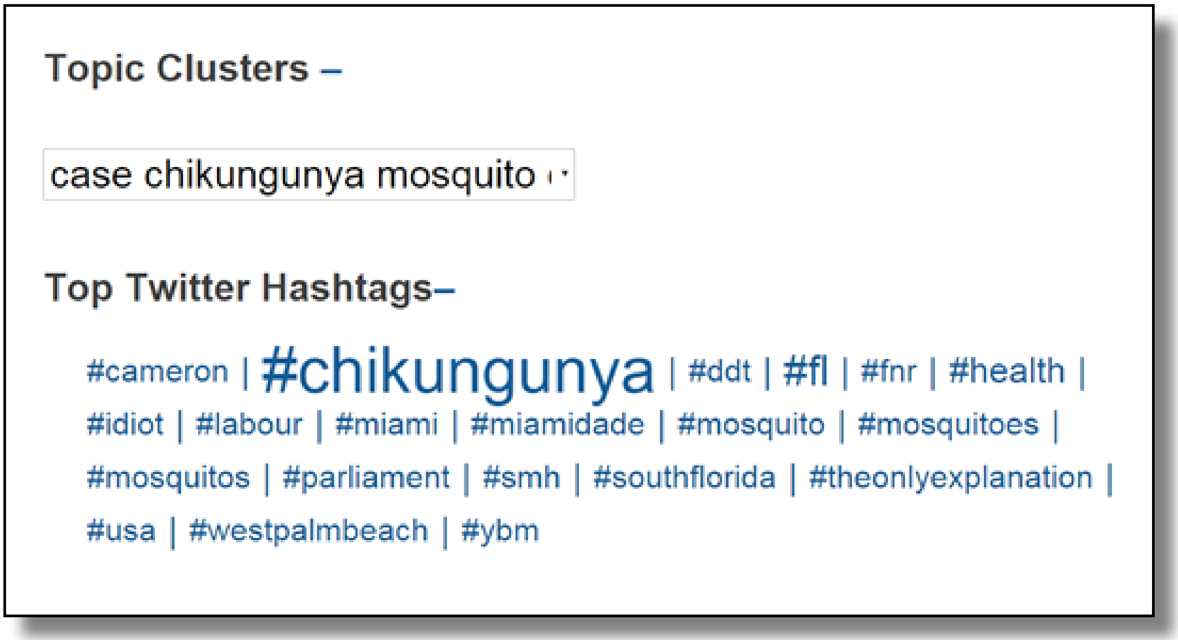
The task of finding relevant biosurveillance information on social media sites is like finding a needle in a haystack. Roughly 500 million tweets are published each day on Twitter (Sayce 2017). According to research performed by the University of Tokyo, 42 percent of the messages on Twitter (tweets) containing a keyword (e.g., "influenza," "Ebola," "H1N1 virus") are false positives, which means that the contents of these tweets are irrelevant to the topics of interest (i.e., influenza, Ebola, or H1N1) (Aramaki, Maskawa, and Morita 2011). The objective of mining social media messages is to reduce the social noise as much as possible to minimize false positives, which can lead to false alarms of an emerging disease.

The other challenge was to discover new topics that arise without prior knowledge, e.g., a new virus breakout or a natural disaster. We employed IDATA's capabilities to address this problem. Although IDATA was not originally intended for social media analysis, it is designed to be highly customizable and extensible. In this case, IDATA was easily extended to ingest tweets, extract hashtags from those tweets, and display the resultant trends.

In 2014, IDA began to feed the first set of tweets into IDATA. To narrow down the scope of the topics, we limited the search to messages that contained a set of keywords related to health, such as "fever," "flu," "influenza," "virus," "infection," "measles," "H1N1," and "pneumonia." IDATA key phrase extraction quickly revealed a high degree of false positives generated by health-related keywords such as "World Cup" (e.g., "World Cup fever") and "Brazil." This illustrates the challenge: A search for "fever" led to posts about the World Cup. We had to determine the best way to use IDATA features to adjust to humans' semantic ambiguity – that is, humans' tendency to ascribe meaning and purpose to words that may differ from the words' original meaning and context (such as *fever* meaning *high temperature* versus *World Cup fever*).

Results

IDATA's underlying analytics, such as topic discovery and entity extraction, in addition to the interactive interface, helped sift through the noise to zero in on buried signals. For example, as shown in Figure 1, we discovered an emerging health topic related to the mosquito-



Topic Clusters –

case chikungunya mosquito

Top Twitter Hashtags–

#cameron | #chikungunya | #ddt | #fl | #fnr | #health | #idiot | #labour | #miami | #miamidade | #mosquito | #mosquitoes | #mosquitos | #parliament | #smh | #southflorida | #theonlyexplanation | #usa | #westpalmbeach | #ybm

Figure 1. IDATA Discovers a New Topic: Chikungunya

borne Chikungunya virus because an individual in Florida was reported to be infected at the time.

IDATA was not originally designed for social media analysis, and its algorithms (e.g., topic models) are not specifically optimized for microblog data, which contain emoticons and a shorthand, colloquial language style. Despite this, IDATA provided good results on this task when applied to Twitter data. Moreover, IDATA's high degree of extensibility made it easy to customize for social media.

References

Aramaki, E., Maskawa, S., and Morita, M. (2011). "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." *Proceedings from the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, July 27-31, 2011.

Defense Threat Reduction Agency. (2014). "The Biosurveillance Ecosystem (BSVE)." http://www.dtra.mil/Portals/61/Documents/bsve-fact-sheet_draft_05-01-2014_pa-cleared-distro-statement.pdf.

Doan, S., Ho Vo, B., and Collier, N. (2011). *An Analysis of Twitter Messages in the 2011 Tohoku Earthquake*. Paper presented at the eHealth 2011 conference, Malaga, Spain.

Draxler, B. (2014). "How Tweets Can Save Lives." *Popular Science*. September 18, 2014 https://www.popsci.com/article/how-tweets-can-save-lives.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors." *Proceedings from the 19th International Conference on World Wide Web*, Raleigh, North Carolina, April 26-30, 2010: 851-860.

Sayce, D. (2017). "Number of Tweets per Day?" *David Sayce*.
https://www.dsayce.com/social-media/tweets-day/.