Disease Outbreak Alert and Characterization using Big Data

Wei Chen, Amy Kircher, DrPH

College of Science & Engineering
Food Protection & Defense Institute

Data Provision

Motivation

Collecting news media data using web crawler

Developing an outbreak ontology

Infectious disease outbreaks can be defined by a relevant set of concepts referred to as classes in an outbreak ontology. These classes include event, time, place, syndrome, epidemiology, control, response, and precursor. Each class can be further defined by relevant subcategories. For example, the response class could include at least three subclasses: response_Local, response_National, and response_International.

Subclass ➔ Class

Event | Time | Place | Syndrome | Response
---|---|---|---|---
Part | Observed | Geography | Symptoms | Local
Whole | First Case | Demography | Signs | National
Last Case | Endemic Diseases | International

Epidemiology | Control | Precursor
---|---|---
Agent | Healthcare Facilities | Environmental
Previously Observed | Water Purification | Weather
Attack Rate | Food Handling | Social Behavior
Case Fatality Rate | Vector Population | Introduced
Case Count | Social Behavior | Political Unrest
Death Count | Trade Restrictions | Economic Difficulties
Mode of Spread | Travel Restrictions | Mode of Spread
Affected Population | Pharmacologic | Affected Population

Data Processing

Natural language processing

Since the words and phrases in the outbreak ontology may not appear exactly the same in the news, natural language processing was used to make subtle variations to words and phrases, such as:
- Stem the words
- Pluralize the nouns
- Change the tense of verbs.

Machine learning for binary classification

The purpose of scraping news from the different sources of media by web crawler is to obtain enough data (>1000 articles) as far as machine learning concerns.

Then, machine learning can be used to train the classifier to identify emerging and expanding disease outbreaks with confidence and to output predictive words and phrases as potential disease signals.

<table>
<thead>
<tr>
<th>Machine Learning Model</th>
<th>Test Set Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial_NB</td>
<td>87.38%</td>
</tr>
<tr>
<td>Bernoulli_NB</td>
<td>83.06%</td>
</tr>
<tr>
<td>Logistic_Regression</td>
<td>89.03%</td>
</tr>
<tr>
<td>SGD_Classifier</td>
<td>86.38%</td>
</tr>
<tr>
<td>LinearSVC_Classifier</td>
<td>84.71%</td>
</tr>
</tbody>
</table>

- The accuracy of classification is around 80% under multiple machine learning algorithms which also return predictive words like 'h1n1' and 'influenza'.

Public Health Analysts' Interaction with system

- Public health analysts examined the machine learning classification results with domain knowledge.
- Inspected the highlighted predictive keywords occurred in the news article.
- Extracted the outbreak properties which were saved in the disease property dictionary.

This research was supported by Paul G. Allen Family Foundation.
Thank you Erin Mann, Rui Kuang and DIT, Inc.