A Real-Time Disease Surveillance Architecture Using Social Networks

Mustafa Sofeare and Matthew Smith
University of Hannover
sofean@dcsec.uni-hannover.de, smith@dcsec.uni-hannover.de

The European Medical Informatics Conference MIE2012
Outline

- The Nature of the Tweet.
- Twitter with public Health.
- Goals.
- Approaches.
- Gold Standard Data.
- Architecture.
- Implementation.
- Evaluation.
The Nature of the Tweet

What's Happening in The World Right Now?

- **Twitter**
  
  A social network site that allows hundreds of millions of users to communicate with each other in real-time.

- **Microblogging**
  
  Twitter is categorized as a Microblogging service which allows users to exchange short sentences up to 140 characters.

- **Real-Time Property.**
  
  Allows actions in the real world to be received unofficially, even before they are reported through official channels.
Twitter with public Health

- Twitter contains a lot of information about diseases.
- Probably good for case counting.
Our Goals

- Tracking the status updates of people with meta data (*location, time stamping*) in *real-time* and identify if it is disease-reporting.

- Constructing a global framework for detecting of symptoms, diseases and monitoring outbreaks using social networks.
  - Collecting and filtering.
  - Named entity recognition and events extraction.
  - Case *counting* and event *alerting*.
Our Approaches

- We use Twitter that offers opportunities for disease surveillance.
- We exploited the public health information that coming directly from the population in real-time.
- Using data mining techniques synchronically to index, extract and cleansing disease-related postings.
- Using state of the art text classification to filter the pure disease-related postings.
Gold Standard Data

- Build a classifier that allows to automatically distinguish tweets that are relevant for disease surveillance and those that are irrelevant.
- Adopted well guidelines for defining clinical cases to build a dataset for a disease-reporting classifier.
- The dataset contain tweets about viruses, bacteriums, normal diseases and symptoms.

**Dataset**

<table>
<thead>
<tr>
<th>Positive disease-related tweets</th>
<th>2380</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive disease-related tweets</td>
<td>3795</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6175</strong></td>
</tr>
</tbody>
</table>

**Model:** Supervised machine learning algorithms with 10-fold cross validation.

**Results:** the performance of the classifier achieved up to 88% of accuracy.

Architecture

Uses Twitter streaming API to get status updates filtered by medical keywords

1. Allow efficient, fast and accurate metadata used to control the scalability of the architecture.
2. Indexes social posts with its metadata (e.g.; user location and time stamping).

Cleans undesired tokens from the post (e.g.; twitter user names, URL's, emoticons, special symbols, numeric values and slang words.)

uses the state of the art text classification to filter postings that are relevant to diseases and those are irrelevant.

1. Working as search engine for extracting postings with their metadata from Index-Pool by using medical keywords-related query.
2. Ignoring tweets which: coming from official resources (e.g.; WHO), related to shot or vaccine, Bieber Fever…etc
Architecture

- Five processes are executed by the architecture respectively:

Implementation

- Java under Ubuntu Linux 11.04 operating system.
- **Crawler** is working independent over time and returns data in JSON format each one contains 1000 postings with their metadata.
- **Indexer** and **Extractor** use Apache Lucene to index and extract data respectively.
- **Classifier** uses the state of the art model (SVM) which produced by WEKA tool.
- MySQL are used to create **temporal** and **central** databases.
- **Multithreading technique** are used to make all components working together over time whenever each gets new data.
Evaluation

- Classifier Performance
  - 10-fold cross-validation
    We used 10-fold cross-validation method with support vector machine (SVM) on the whole dataset, we got the performance with accuracy 88.261%.
    - In addition, we collected 200 tweets; all these tweets belong to the public health domain and contain various names diseases. 111 tweets annotated manually as disease-related (Positive) and 89 tweets annotated as non disease-related (Negative). the performance of accuracy for classification model was 85.5%.
Evaluation

- **Run Time Performance**
  
  We run our performance experiment on Intel Xeon 2.40GHZ (8 processors) with 2G of memory; running a 32-bit version of Ubuntu Linux 11.04.

  - **Latency**
  - **Throughput.**
### Evaluation

- Run Time Performance
  - Latency

<table>
<thead>
<tr>
<th>Component</th>
<th>Run Time (Millisecond)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexer</td>
<td>0.487</td>
</tr>
<tr>
<td>Extractor</td>
<td>11</td>
</tr>
<tr>
<td>Classier</td>
<td>0.975</td>
</tr>
<tr>
<td>Total</td>
<td>12.462</td>
</tr>
</tbody>
</table>
Evaluation

- Run Time Performance
  - **Throughput**: We run our system over time for 6 days.
Thank You

Thank you for your attention and tweeting on your personal diseases.

Questions

?