Twitter improves Seasonal Influenza Prediction

- Harshavardhan Achrekar [1]
- Avinash Gandhe [2]
- Ross Lazarus [3]
- Ssu-Hsin Yu [2]
- Benyuan Liu [1]

HEALTHINF 2012
International Conference on Health Informatics
Vilamoura, Algarve, Portugal

[1] Computer Science Department, University of Massachusetts Lowell
[3] Department of Population Medicine – Harvard Medical School
Talk Outline

- Background
- Related Work
- Our Approach
- Twitter Dataset Description
- Twitter Dataset Analysis
- Extended Analysis (Text Mining, Age-wise, US Regions)
- Conclusions
Influenza (flu) is a contagious respiratory illness caused by influenza viruses.

Seasonal - wave occurrence pattern.

5 to 20% of the population gets flu

≈ 200,000 people are hospitalized from flu-related complications.

36,000 people die from flu every year in the USA, worldwide death toll is 250,000 to 500,000.

Epidemiologists want to use early detection of disease outbreak to reduce the number of people affected.

CDC collects Influenza-like Illnesses (ILI) from its surveillance network and publishes weekly (typically 1–2 weeks delay)

CDC stands for Center for Disease Control
Emerging Flu Epidemics

- Spanish Flu
- SARS, 2002-2003
- H1N1 (swine flu), 2009-2010
- ?
Related Work :- Google Flu Trends

- Over the counter drug sale, patients visit log for flu shots, tapping telephone advice lines.
- Certain Web Search terms are good Indicators of flu activity.
- Google Trend uses Aggregated search data on flu indicators.
- Estimate current flu activity around the world in real time.
- From example :- Google Flu Trend detects increased flu activity two weeks before CDC.

Link:- www.google.com/flutrends
OSN - Novel Data Source for Detection and Prediction

Online Social Networks (OSN) has emerged as a popular platform for people to make connections, share information and interact.

- Facebook: ~750 million users, Twitter: ~200 million users
- Billions of pieces of information being posted and shared on the web every week.

Applications:
- Real-world outcome of box-office revenues for movie
- Large scale fire emergencies and Earthquake detection and reporting
- Online Service Downtime and disruptions of content providers
- People’s mood
- Live Traffic updates
Our Approach

• {“i am down with flu”, “got flu.”} msg exchange between users provide early, robust predictions.

• OSN represent a previously untapped data source for detecting onset of an epidemic and predicting its spread.

• Twitter/Facebook mobile users tweet/posts updates with their geo-location updates. helps in carrying out refined analysis.

• User demographics like age, gender, location, affiliated networks, etc can be inferred from data.

• snapshot of current epidemic condition and preview on what to expect next on daily or hourly bases.

• sought to develop model that estimates number of physician visits per week related to ILI as reported by CDC.

OSN stands for Online Social Network
System Architecture of SNEFT

ILI stands for Influenza-Like Illness
OSN Data Collection

Design of the Twitter data collection engine / Crawler
Twitter Data Set

- Real Time Response Stream fetches entries relevant to searched keyword having the tweets in reverse-time order.
- Data collection active from October 18, 2009 until present.
  - 2009–2010: 4.7 million tweets from 1.5 million unique users
  - 2010–2011: 4.5 million tweets from 1.9 million unique users
Crawler uses Streaming Real time Search Application Programming Interface (API) to fetch data at regular time intervals. A tweet has the

- Twitter User Name,
- the Post with status id
- Time stamp attached with each post.

From Twitter’s username we can get profile details attached to every user which include

- number of followers,
- number of friends,
- his/her profile creation date,
- location {public or private from the profile page or mobile client}
- with status updates count

User’s current location is passed as an input to Google’s location based web services to get geo-location codes (i.e., latitude and longitude) along with the country, state, city with a certain accuracy scale.
Twitter Data Set Analysis

- In our Twitter dataset [2010–2011]
  - 22% users are from USA,
  - 46% users are outside USA
  - 32% users have not published their location details.

- Status posting times (tweet timestamp in GMT) are converted to the local timezone of the individual profile. Day light saving are applied within required time frame.

State-wise Distribution of USA users on Twitter for flu postings

Monday, January 30, 2012
Hourly Twitter usage pattern in USA

The hourly activity patterns observed at different hours of the day are much to our expectations, with high traffic volumes being witnessed from late morning to early afternoon and less tweet posted from midnight to early morning, reflecting people’s work and rest hours within a day.

Average daily usage pattern within a week suggests a trend on OSN sites with more people discussing about flu on weekdays than on weekends.
Twitter Data Set Cleaning

- “I got flu shot”, “got stomach flu”, “flu season”...do not indicate real flu events
- Need to classify the tweets into positive or negative categories.
- 25,000 tweets classified by Amazon Mechanical Turk as training set
- Use trained SVM to classify all other tweets
- Significant improvement of correlation between the Twitter data and CDC data.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 decision tree</td>
<td>Yes</td>
<td>0.801</td>
<td>0.791</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.813</td>
<td>0.704</td>
<td>0.755</td>
</tr>
<tr>
<td>Naive Bayesian</td>
<td>Yes</td>
<td>0.725</td>
<td>0.829</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.813</td>
<td>0.704</td>
<td>0.755</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Yes</td>
<td>0.807</td>
<td>0.822</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.829</td>
<td>0.814</td>
<td>0.822</td>
</tr>
</tbody>
</table>

Text Classification 10 fold cross validation results

Classified Twitter dataset achieves higher correlation with CDC reports on Nationwide and Regional levels.
Retweet: A retweet is a post originally made by one user that is forwarded by another user.

Syndrome elapsed time: An individual patient may have multiple encounters associated with a single episode of illness. To avoid duplication the first encounter for each patient within any single syndrome group is reported to CDC, but subsequent encounters with the same syndrome are not reported as new episodes until more than six weeks has elapsed since the most recent encounter in the same syndrome. We call it syndrome elapsed time.

Remove retweets and tweets from the same user within a certain syndrome elapsed time, since they do not indicate new ILI cases.

<table>
<thead>
<tr>
<th>Retweet</th>
<th>Syndrome Elapse Time</th>
<th>Correlation coefficient</th>
<th>RMSE errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0 week</td>
<td>0.8907</td>
<td>0.3796</td>
</tr>
<tr>
<td>No</td>
<td>1 week</td>
<td>0.8895</td>
<td>0.3818</td>
</tr>
<tr>
<td>No</td>
<td>2 week</td>
<td>0.8886</td>
<td>0.3834</td>
</tr>
<tr>
<td>No</td>
<td>3 week</td>
<td>0.886</td>
<td>0.3878</td>
</tr>
<tr>
<td>No</td>
<td>4 week</td>
<td>0.8814</td>
<td>0.3955</td>
</tr>
</tbody>
</table>

Correlation between Twitter dataset and CDC along with Root Mean Square Errors (RMSE).
Twitter Data Set Analysis

Data show strong correlation (Pearson correlation coefficient 0.8907) between Twitter data set and ILI rates from CDC, providing a strong base for accurate prediction of ILI rate.
Prediction Model

Auto Regressive model with external input (Twitter data)

Logistic ARX Model

\[
\log \left( \frac{y(t)}{1 - y(t)} \right) = \sum_{i=1}^{m} a_i \log \left( \frac{y(t-i)}{1 - y(t-i)} \right) + \sum_{j=0}^{n-1} b_j \log(u(t-j)) + c + e(t)
\]

- \( t \): indexes weeks
- \( y(t) \): the percentage of physician visits due to ILI in week \( t \)
- \( u(t) \): the number of unique Twitter users with flu related tweets in week \( t \)
- \( e(t) \): is a sequence of independent random variables
- \( c \): is a constant term to account for offset.
- \( m \): previous CDC data in weeks
- \( n \): previous Twitter data in weeks
### Cross Validation Results

<table>
<thead>
<tr>
<th></th>
<th>n=0</th>
<th>n=1</th>
<th>n=2</th>
<th>n=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>m=0</td>
<td>0.5355</td>
<td>0.4814</td>
<td>0.4813</td>
<td></td>
</tr>
<tr>
<td>m=1</td>
<td>0.6331</td>
<td>0.4107</td>
<td>0.4147</td>
<td>0.4314</td>
</tr>
<tr>
<td>m=2</td>
<td>0.5395</td>
<td>0.3957</td>
<td>0.3986</td>
<td>0.4256</td>
</tr>
</tbody>
</table>

#### Root Mean Squared Errors from 10-fold cross validation

Weekly plot of percentage of weighted ILI visits, positively classified Twitter dataset and predicted ILI rate using CDC and Twitter

- Addition of Twitter data improves the prediction with past CDC data alone.
- Use of Twitter data alone to predict the ILI rate (m=0) results in poor predictions.
- Best result when m=2, n=1: previous 2 week’s CDC data, current Twitter data.
Regional and Age-based Flu Prediction Analysis

Health and Human Services Regions
ILI seems to peak later in the northeast (Region 1 and 2) than in the rest of the country by at least week. The Twitter reports also follow this trend.

In Region 9 (CA, NV, AZ...), Region 4 (FL, etc.) and the northeast, the ILI rates seem to drop off fairly slowly in the weeks immediately following the peaks. This is also reflected in the Twitter reports.

Approximately 20–25 weeks after the peak ILI, the northern regions have lower levels relative to the peaks in the southern regions. This is true for Twitter reports.

<table>
<thead>
<tr>
<th>Week</th>
<th>CDC ILI data</th>
<th>CDC week</th>
<th>Twitter Reports</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009 Week 43</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td>Both the ILI and Twitter numbers are at their maximum except in the northeast.</td>
</tr>
<tr>
<td>2009 Week 44</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td>The ILI and Twitter numbers both peak in northeast during this week. The drop in ILI in Regions 2 and 9 is small, which is also reflected in the Twitter numbers.</td>
</tr>
<tr>
<td>2009 Week 45</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td>Areas that show small drop-off in ILI rates (Regions 1, 2, 4, 10) also show small drops in Twitter numbers. The mid-western states show large drops in both ILI and Twitter numbers.</td>
</tr>
<tr>
<td>2009 Week 47</td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td>The ILI incidence in the Southern US remains significant and this reflected to some extent in the Twitter numbers.</td>
</tr>
<tr>
<td>2009 Week 50</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>
Twitter data is a good indicator of ILI rates and can be used to effectively improve the regional prediction of current ILI rates.
The results indicates that for most of the regions, Twitter data best fits the age-groups of 5–24 yrs and 25–49 yrs, which correlates well with the fact that this likely is the most active age groups using Twitter.

Prediction performance (root relative squared error) using Twitter in different age groups for different geographical regions within the US.
Conclusions

- Investigated use of a previously untapped data source, namely, messages posted on Twitter to track and predict influenza epidemic situation in real world.
- Results show that the number of flu related tweets are highly correlated with ILI activity in CDC data (Pearson correlation coefficient 0.8907).
- Build logistic auto-regression models to predict number of ILI cases in a population as percentage of visits to physicians in successive weeks.
- Verified that Twitter data substantially improves our model’s accuracy in predicting ILI cases nationwide as well as region-wise.
- Twitter best fits age group 5–24 and 25–49 years, as these are most active age group communities on Twitter.
- Opportunity to significantly enhance public health preparedness among the masses for influenza epidemic and other large scale pandemic.

Thank You

Presenter: Harshavardhan Achrekar
Personal Website: www.hdachrekar.com
CDC’s Region-wise ILI data (left) and Twitter data (right)