User Level Sentiment Analysis Incorporating Social Networks

Chenhao Tan

Department of Computer Science
Cornell University

Joint work with:
Lillian Lee, Jie Tang, Long Jiang, Ming Zhou and Ping Li

May 18, 2011
Outline

1 Motivation
2 Problem Setting in Twitter
3 Data Collection
4 Observation
5 Model
6 Approach
7 Experiment
8 Conclusion
Motivation

- User-level sentiment analysis
- Network information
  - Accessibility
  - Homophily or Attention
Twitter as the basis

- Text information: Tweets
- Network information
  - Follow Network
  - @ Network
  - directed
  - mutual
Semi-supervised Learning in Twitter

- Hard to get full labels
- Given a graph and labels of some nodes in the graph, try to classify the other users in the graph
Data Collection

- Traditional Annotation by Tweets
Failed Traditional Annotation by Tweets
Data Collection

- Failed Traditional Annotation by Tweets
- User Biographical Information
Final Data Set

- 1,414,340 users
- 1,414,211 user profiles
- 480,435,500 tweets
- 274,644,047 t-follow edges
- 58,387,964 @-edges

<table>
<thead>
<tr>
<th>Topic</th>
<th>#labeled users</th>
<th>#t-follow edges (mutual)</th>
<th>#@ edges (mutual)</th>
<th>#users in total</th>
<th>total # of on-topic tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>889</td>
<td>7838(2949)</td>
<td>2358(302)</td>
<td>307985</td>
<td>4873711</td>
</tr>
<tr>
<td>Sarah Palin</td>
<td>310</td>
<td>1003(264)</td>
<td>449(60)</td>
<td>121910</td>
<td>972537</td>
</tr>
<tr>
<td>Glenn Beck</td>
<td>313</td>
<td>486(159)</td>
<td>148(17)</td>
<td>95847</td>
<td>687913</td>
</tr>
<tr>
<td>Lakers</td>
<td>640</td>
<td>2297(353)</td>
<td>1167(127)</td>
<td>76926</td>
<td>301558</td>
</tr>
<tr>
<td>Fox News</td>
<td>231</td>
<td>130(32)</td>
<td>37(5)</td>
<td>114530</td>
<td>1231519</td>
</tr>
</tbody>
</table>
Sharing Label conditioned on being connected

- Probability that two users have the same label, conditioned on whether or not they are connected

### t-Follow Graph

- random
- directed
- mutual

### @ Graph

- random
- directed
- mutual

Chenhao Tan
Microsoft Research Asia
Connectedness conditioned on labels

- Probability that two users are connected, conditioned on whether or not they have the same label
Model Framework

- **User-Tweet Factor**

\[
f_{k,l}(y_i, y_t) = \begin{cases} 
\frac{w_{\text{labeled}}}{|\text{tweet}_{v_i}|} & y_i = k, y_t = l \\
\frac{w_{\text{unlabeled}}}{|\text{tweet}_{v_i}|} & y_i = k, y_t = l \\
0 & \text{otherwise}
\end{cases}
\]

- **User-User Factor**

\[
h_{k,l}(y_i, y_j) = \begin{cases} 
\frac{w_{\text{relation}}}{|\text{Neighbors}_{v_i}|} & y_i = k, y_j = l \\
0 & \text{otherwise}
\end{cases}
\]

- **Objective Function**

\[
\log P(Y) = \left( \sum_{v_i \in V} \left[ \sum_{t \in \text{tweet}_{v_i}, k, l} \mu_{k,l} f_{k,l}(y_{v_i}, y_t) + \sum_{v_j \in \text{Neighbors}_{v_i}, k, l} \lambda_{k,l} h_{k,l}(y_{v_i}, y_{v_j}) \right] \right) - \log Z
\]
Approach

- Parameter Estimation
  - Direct estimation from simple statistics
  - SampleRank
- Inference
  - loopy belief propagation
Methods

- **Training set:** 50 positive users and 50 negative users, the others for testing
- **Labels of Tweets**
  - SpecificSVM
- **Labels of Users**
  - Majority Vote
  - HGM-NoLearning
  - HGM-Learning
Case Study

(a) Ground Truth  (b) Text-Only Approach  (c) Our algorithm
Sample tweets of users classified correctly only with network information

<table>
<thead>
<tr>
<th>User ID</th>
<th>SVM</th>
<th>HGM</th>
<th>True</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NEG</td>
<td>POS</td>
<td>POS</td>
<td>Obama is making the repubs look silly and petty. #hrc</td>
</tr>
<tr>
<td>2</td>
<td>NEG</td>
<td>POS</td>
<td>POS</td>
<td>Is happy Obama is President</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Tweet link]</td>
</tr>
<tr>
<td>3</td>
<td>NEG</td>
<td>POS</td>
<td>POS</td>
<td>I am praying that the government is able to get health care reformed this year! President Obama seems like the ONE to get it worked out!!</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Watching House on TV. I will be turning to watch Rachel M. next. I am hoping Pres. Obama gets his budget passed. Especially Health Care!</td>
</tr>
<tr>
<td>4</td>
<td>POS</td>
<td>NEG</td>
<td>NEG</td>
<td>RT @TeaPartyProtest Only thing we have 2 fear is Obama himself &amp; Pelosi &amp; Cong &amp; liberal news &amp; Dems &amp;... [Tweet link]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RT @GlennBeckClips: Barack Obama can no more disown ACORN than he could disown his own grandmother. #TCOT</td>
</tr>
<tr>
<td>5</td>
<td>POS</td>
<td>NEG</td>
<td>NEG</td>
<td>RT @JosephAGallant Twitlonger: Suppose I wanted to Immigrant to Mexico? A Letter to President Obama... [Tweet link]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>George Bush was and acted like a war time President. Obama is on a four year power grab and photo op. #icot</td>
</tr>
<tr>
<td>6</td>
<td>POS</td>
<td>NEG</td>
<td>NEG</td>
<td>ObamaCare forces Americans to buy or face a fine! It is UNCONSTITUTIONAL to force us to buy obamacare. Marxist Govt. taking our Freedoms!</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Look up Chicago Climate Exchange, an organization formed years ago by Obama &amp; his Marxist-Commie Cronies to form a profit off cap &amp; trade.</td>
</tr>
</tbody>
</table>
**Overall Performance**

- Beat Baseline!
- *Follow* better than *@*
- *Directed* better than *Undirected*
- NoLearning same with Learning
- Sparseness of graph
- Size of graph or #Tweets per user
- SVM Classifier Performance
Adding More Unlabeled Data

- Learning better than NoLearning
Empirical analyses on the correlation of networks and sentiment
Propose a heterogeneous graphical model
Validate the effectiveness of incorporating network information
Future Work

- More data sets
- Better models and semi-supervised learning algorithms
- Find the helpful parts of networks
- Build a theory of why and how users correlate on different topics in different kinds of networks
The End

Thank you!
Questions?