Mining Twitter to Aid Disaster Response

Doina Caragea
Computer Science
Kansas State University

Elizabeth Unger Celebration - April 29, 2016
Acknowledgements

HongMin Li, KSU
Cornelia Caragea, UNT
Nic Herndon, KSU
Andrea Tapia, PSU
Twitter contains valuable information
How can one identify useful content?

- Possible approaches:
  - Keyword search: noisy, non-relevant tweets retrieved
  - Manual selection: too time consuming
  - Supervised learning: a large number of labeled tweets needed [Imran et al., 2013; Ashktorab et al., 2014; Caragea et al., 2014]

- Our approach [Li et al, 2015]:
  - Domain adaptation using labeled data from a previous source disaster and unlabeled data from the current target disaster
Domain Adaptation

Challenge: distributions of source domain and target domain are generally different

- Need to identify knowledge that can be transferred from source to target
- Need to identify approaches that can be used to transfer knowledge

Our approach to transferring knowledge:

- Naïve Bayes for Domain Adaptation [Herndon and Caragea, 2013; 2014] inspired from [Tan et al., 2009]
- An iterative approach of expectation maximization (EM)
Naïve Bayes for Domain Adaptation

1. Compute the prior and likelihood based on labeled source data, using only generalizable features from the source

\[ P(c_k) = P_S(c_k) \]
\[ P(w_i \mid c_k) = P_S(w_i \mid c_k) \]
2. Assign labels to the unlabeled tweets from the target

\[ P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k) \]
Naïve Bayes for Domain Adaptation

3. While labels assigned to unlabeled tweets change
   - **M-step**: compute prior and likelihood based on source and newly labeled target tweets

\[
P(c_k) = (1 - \lambda)P_S(c_k) + \lambda P_T(c_k)
\]

\[
P(w_i \mid c_k) = (1 - \lambda)P_S(w_i \mid c_k) + \lambda P_T(w_i \mid c_k)
\]
3. While labels assigned to unlabeled tweets change
   - **E-step**: Assign labels to the unlabeled tweets from the target domain

\[ P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k) \]
Naïve Bayes for Domain Adaptation

4. Use the final classifier to assign labels to the test target tweets

\[ P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k) \]
Research Questions:

• Does domain adaptation help?
  • Compare supervised Naïve Bayes classifiers with domain adaptation Naïve Bayes classifiers

• How much labeled source data is needed for accurate target classification?
  • Vary the amount of source labeled data

• Does the distance between the domains affect the classification accuracy?
  • Compare results for groups of more similar disasters with results for groups of more different disasters
### Datasets

CrisisLexT6 [Olteanu et al., 2014]

<table>
<thead>
<tr>
<th>Disaster event</th>
<th>On-topic</th>
<th>Off-topic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012_Sandy_Hurricane</td>
<td>5,261</td>
<td>3,752</td>
<td>9,013</td>
</tr>
<tr>
<td>2013_Queensland_Floods</td>
<td>3,236</td>
<td>4,550</td>
<td>7,786</td>
</tr>
<tr>
<td>2013_Boston_Bombings</td>
<td>4,441</td>
<td>4,309</td>
<td>8,750</td>
</tr>
<tr>
<td>2013_Alberta_Floods</td>
<td>3,497</td>
<td>4,714</td>
<td>8,211</td>
</tr>
</tbody>
</table>
Disaster groups chosen for experiments

<table>
<thead>
<tr>
<th>Group</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>2013_Queensland_Floods</td>
<td>2013_Alberta_Floods</td>
</tr>
<tr>
<td>Group 2</td>
<td>2012_Sandy_Hurricane</td>
<td>2013_Alberta_Floods</td>
</tr>
<tr>
<td>Group 3</td>
<td>2013_Boston_Bombing</td>
<td>2013_Alberta_Floods</td>
</tr>
</tbody>
</table>
Experimental Results: Group 1

Source: 2013_Queensland_Floods
Target: 2013_Alberta_Floods

- NB
- NB_DA
Experimental Results: Group 2

Source: 2012_Sandy_Hurricane
Target: 2013_Alberta_Floods
Experimental Results: Group 3

Source: 2013_Boston_Bombing
Target: 2013_Alberta_Floods
Conclusions

- Domain adaptation helps improve the classification accuracy as compared to supervised learning.
- Source data is useful even in small amounts (e.g., 500 instances).
- Larger amount of source labeled data result in better accuracy.
- The perceived similarity between source and target does not always give better results.
Future Work

• Extend the algorithm from single source to multi source
• Identify importance weights for sources automatically (e.g., using Kullback-Leibler divergence)
• Identify useful instances in each source and remove those that are not useful