Latent Attribute Detection in Social Media

Glen Coppersmith, Clay Fink, Tim Oates, Michael Paul, Delip Rao, David Yarowsky

Thursday, February 3, 2011
Nigeria Users in Social Media
Nigeria Users in Social Media

- 39,574 Users
- 13.8M Tweets
- 4.5M conversations
Nigeria Users in Social Media

- 39,574 Users
- 13.8M Tweets
- 4.5M conversations

- 23,848 Users
- 68,772 conversations
- Full name
- Self-identified gender
Nigeria Users in Social Media

- 39,574 Users
- 13.8M Tweets
- 4.5M conversations

Goal: Learn latent attributes like gender and ethnicity from Tweet/Status messages and/or network structure.

- 23,848 Users
- 68,772 conversations
- Full name
- Self-identified gender

Thursday, February 3, 2011
Additional Data

- Scrape name dictionaries
  - 1980 names
  - Lack full names
Additional Data

- Scrape name dictionaries
  - 1980 names
  - Lack full names
Additional Data

• Scrape name dictionaries
  – 1980 names
  – Lack full names
MT Turk Annotations

Identify the Nigerian ethnic group associated with each of these names.

Collins Audu Difa

- Hausa
- Igbo
- Yoruba
- Other Nigerian Ethnicity (Please note)
- Non-Nigerian Ethnicity (Please note)
- Unknown

Note (Optional)
Name-based models

• How useful are names?
• Twitter – Unreliable
  – “yoyo”
  – “Midnight Crew”
• Facebook – Mostly reliable
From Names to Features

abasifrek

e
From Names to Features

abasifrek
e
aba
bas
asi
sif
ifr
fre
rek
ake
From Names to Features

abasifrek e

aba–BEG
bas–MID
asi–MID
sif–MID
ifr–MID
fre–MID
rek–MID
ekte–END
From Names to Features

- {2,3,4,5}-grams
- WORD features
- Positional features
- Prefix/Suffix
## Name Model: Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>50.0</td>
<td>40.8</td>
</tr>
<tr>
<td>SVM</td>
<td>78.4</td>
<td>87.8</td>
</tr>
<tr>
<td>Maxent</td>
<td>80.7</td>
<td>86.3</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>77.7</td>
<td>87.8</td>
</tr>
<tr>
<td>UNHB</td>
<td>77.5</td>
<td>90.0</td>
</tr>
</tbody>
</table>
User Name Hierarchical Bayes (UNHB)

Chang et al (2010):

\[ \delta \rightarrow \eta \rightarrow \sigma \rightarrow \psi_f \rightarrow a \rightarrow m_f \rightarrow m_1 \rightarrow \psi_1 \]

Our model:

\[ \delta \rightarrow \eta \rightarrow \sigma \rightarrow \psi \rightarrow a \rightarrow m \rightarrow F \rightarrow U \]
User Name Hierarchical Bayes (UNHB)

1. For each user, sample an attribute \( a \) from a distribution over the aggregate population

2. Sample name features from \( p(\text{name}|a) \)
   - The conditional distributions over names may be known a priori
   - Chang et al. (2010) used U.S. Census data to define a distribution over surnames conditioned on ethnicity
Content-based Models

- Extract features from status messages
  - \{1,2\}-gram features
  - Sociolinguistic features

Rao & Yarowsky, 2010 in NIPS Workshop on Machine Learning for Social Networks

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>Description/Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMLEYS</td>
<td>A list of emoticons compiled from the Wikipedia. Abbreviation for ‘Oh My God’</td>
</tr>
<tr>
<td>OMG</td>
<td>‘....’</td>
</tr>
<tr>
<td>ELLIPSES</td>
<td>E.g. my_XXX, our_XXX</td>
</tr>
<tr>
<td>POSSESSIVE BIGRAMS</td>
<td>E.g. niceeeeee, noooo waaaay</td>
</tr>
<tr>
<td>REPATED ALPHABETS</td>
<td>E.g., I_XXX, I’m_XXX</td>
</tr>
<tr>
<td>SELF</td>
<td>E.g. LOL, ROTFL, LMFAO, haha, hehe</td>
</tr>
<tr>
<td>LAUGH</td>
<td>Text in ALLCAPS</td>
</tr>
<tr>
<td>SHOUT</td>
<td>E.g. Ugh, mmmm, hmmm, ahh, grrr</td>
</tr>
<tr>
<td>EXASUREATION</td>
<td>E.g. yea, yeah, ohya</td>
</tr>
<tr>
<td>AGREEMENT</td>
<td>E.g. dude, man, bro, sir</td>
</tr>
<tr>
<td>HONORIFICS</td>
<td>E.g. xoxo</td>
</tr>
<tr>
<td>AFFECTION</td>
<td>A string of exclamation symbols (!!!!!)</td>
</tr>
<tr>
<td>EXCITEMENT</td>
<td>A single exclamation at the end of the tweet</td>
</tr>
<tr>
<td>SINGLE EXCLAIM</td>
<td>A combination of any number of ? and ! (!?!!??!)</td>
</tr>
<tr>
<td>PUZZLED PUNCT</td>
<td></td>
</tr>
</tbody>
</table>
Topic Modeling

• Latent Dirichlet Allocation (LDA)
• Associates each feature in the corpus with a latent “topic” variable
• Provides a model for representing data in a lower dimensional space
• Useful for clustering

1. Draw a distribution over topics
   \( \theta \sim \text{Dirichlet}(\alpha) \)
2. For each token in the document, sample a topic \( \sim \theta \) and then sample a feature from \( P(\text{feature}|\text{topic}) \)
Topic Model Variants

• Single-user model
  – 1 document = 1 message from a user
  – 1 document = all messages from a user

• Dyadic model
  – 1 document = 1 conversation between two users
  – May be useful for analyzing relations and interactions between people
# Single-User Model: Example

<table>
<thead>
<tr>
<th>Male Users</th>
<th>Female Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUDE</td>
<td>SMILIE</td>
</tr>
<tr>
<td>bro</td>
<td>SMILIE_&lt;3</td>
</tr>
<tr>
<td>man</td>
<td>REPEATED_CHAR</td>
</tr>
<tr>
<td>bros</td>
<td>SMILIE_:)</td>
</tr>
<tr>
<td>tanx</td>
<td>DIGITS_DIGITS</td>
</tr>
<tr>
<td>amen</td>
<td>ELIPSES</td>
</tr>
<tr>
<td>my_man</td>
<td>aku</td>
</tr>
</tbody>
</table>
## Dyadic Model: Example

### Conversation Topics

<table>
<thead>
<tr>
<th>miss</th>
<th>wats up</th>
</tr>
</thead>
<tbody>
<tr>
<td>miss_u</td>
<td>what_s</td>
</tr>
<tr>
<td>yhu</td>
<td>x</td>
</tr>
<tr>
<td>i_miss</td>
<td>wats_up</td>
</tr>
<tr>
<td>missed</td>
<td>gud</td>
</tr>
<tr>
<td>boo</td>
<td>hey</td>
</tr>
<tr>
<td>u_too</td>
<td>s_up</td>
</tr>
</tbody>
</table>
Combining unlabeled data using Topics

\[ P(a \mid u) = \sum_{t} P(a \mid t)P(t \mid u) \]
Community Detection

• Users in a community share attributes
• Multiple ways to define communities
  – Communication
  – Topical
  – Geographical
  – ...

• Look at communication graph for
Community Detection

• Users in a community share attributes
• Multiple ways to define communities
  – Communication
  – Topical
  – Geographical
  – ...

• Look at communication graph for

\[ P(a \mid u) = \sum_t P(a \mid t)P(t \mid u) \]
Community Detection

• Users in a community share attributes
• Multiple ways to define communities
  – Communication
  – Topical
  – Geographical
  – ...

\[ P(a \mid u) = \sum_t P(a \mid t)P(t \mid u) \]

\[ P(t \mid \text{community}) \approx P(t \mid u) \]

• Look at communication graph for
Ethnicity Encoded in Context
Ethnicity Encoded in Context
Ethnicity Encoded in Context
Labels from 2-Neighborhood

Density

Proportion Yoruba
(N=7185 of 11431 Usernames)

Label Proportion

Margin Size

Thursday, February 3, 2011
Labels from Girvan &
Combining Names and
Combined Name+Topic

1. For each user, sample an attribute $a$ from a distribution over the aggregate population
2. Sample name features from $p(name|a)$
3. Draw a distribution over user’s topics $\theta \sim \text{Dirichlet}(\alpha_a)$
4. For each token in the user’s messages, sample a topic from $\theta$ and then sample a word based on the topic
Name+Content: Results

• Sample of the Facebook data
  – 2444 users
  – 534,096 content features
  – 145,595 name features

<table>
<thead>
<tr>
<th></th>
<th>Prior</th>
<th>Unsupervised</th>
<th>Semi-supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names only</td>
<td>58.57</td>
<td>73.2</td>
<td>78.6</td>
</tr>
<tr>
<td>Names+content</td>
<td>72.0</td>
<td>79.3</td>
<td></td>
</tr>
</tbody>
</table>
Summary

• Learning Latent Attributes in Social Media
  – Variety of resources (Twitter, Facebook, name dictionaries, MTurk, …)
  – Utility of names, content, context, and combinations
  – Various baselines, semi-supervised, hierarchical Bayes (need to tune parameters)
Questions