Some stuff on identities and networks and stereotypes and text

Kenny Joseph
kjoseph@cs.cmu.edu

Identities are the words and phrases we use to label other people
Stereotypes are the meanings conveyed by an identity

An EPA profile is a position in this 3D space

Affective stereotypes are defined by how we feel about identities

(Osgood, 1969; Heise 1987)
Semantic stereotypes refer to relationships we presume between identities.

Our identities and the stereotypes they carry have important effects on our lives.
Overview

- Extracting affective stereotypes using “social event networks”
- Extracting a network model of stereotypes
- Networks of identities

The Data

- Newspaper data
  - 600K articles
  - LexisNexis, centered on 16 MENA countries
  - Major news outlets
  - 7/10 – 12/12
Measuring Stereotypes with ACT

(Osgood, 1969)

- An affective, attributional measurement model

Inferring Stereotypes using ACT

? officials
- criticize
+ women

ACT gives a mathematical model for how social events imply stereotypes
Caveat to applying event model

?- officials
- accused
?= protestors

Soln. – allow stereotypes to “diffuse”

?- officials
- criticize
+ women
= officials
- accused
+ protestors
A caveat to applying event model

Sometimes, officials ≠ officials

Solution:
Assume multiple latent stereotypes of each identity/behavior exist.

An overview of the approach

Last week, Egyptian officials shot protesters.
More on extracting events, identities

1. Ran dependency parser, extracted all N -> V -> N
2. Cleaned text using, e.g., stemming (accused -> accuse)
3. Hand-curated list of identities and behaviors

- 102 identities, 87 behaviors, 10K events
- Only 44% of identities in ACT dicts

The Statistical Model

\[
\begin{align*}
\pi & \sim \text{Dirichlet}(p) \\
\phi & \sim \text{Dirichlet}(\psi) \\
\theta & \sim \text{Dirichlet}(\alpha) \\
\mu_0 & \sim N(m_0, k_0) \\
\sigma^2_0 & \sim \text{Inv-Chi}^2(v_0, s_0) \\
a & \sim \text{Categorical}(\theta) \\
o & \sim \text{Categorical}(\theta) \\
b & \sim \text{Categorical}(Q)
\end{align*}
\]

![Diagram of the Statistical Model](image)
One Result w.r.t. religious identities

- Sunnis universally bad, powerful
- Explanation:
  - Events on the ground
  - Western media bias?

NETWORK MODELS OF STEREOTYPE
Parallel Constraint Satisfaction Models

Links in PCSMs define semantic stereotypes

PCSMs are essentially Markov Random Fields through which cognitive activation flows

Hard to model Affect in PCSMs

?
Combining existing models

Attributional
Parsimonious, Affective
No semantic relationships

Relational
Cognitively More Plausible, Semantics
No affective meaning

Affective + Semantic Network of Stereotypes

Stereotypes as an attributed network
Now, how do we “learn” from Twitter data?
Data Used (Population considered)

- Twitter data
  - Subset of 50K users from Study 2
  - Subsetting based on more restricted bot/celeb removal, gender tagging (gender not used)

- 310 identities of interest
  - From popular identities in Study 2 results; some domain relevant

- Sentiment data (EPA profiles)
  - Smith-Lovin et al. (2015)
  - Warriner et al. (2014)

A Statistical Model for Twitter Data

Semantic Model

Affective Model
(Survey data priors)

Tweet data (per user)
Generative model – affective stereotypes

\[
p(\mu, \Sigma) \sim \mathcal{N}I\mathcal{W}(\mu_0, \Sigma_0, \kappa_{0,S}, \gamma_{0,S})
\]

\[
p(\phi) \sim \mathcal{N}(\mu, \Sigma)
\]

\[
p(d) \sim \text{Laplace}(q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z), \beta)
\]

- Draw per-identity distrib. in EPA space from survey priors
- Draw per-user EPA profiles from this distribution
- Draw per-tweet "deflection" balancing by user’s current views, constraints in tweet

Details on deflection

\[
p(d) \sim \text{Laplace}(q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z), \beta)
\]

- In ACT, deflection defines likelihood of social event
  - “Teacher instructs student” has low deflection
  - “Teacher hits student” has high deflection
- I use the same concept for likelihood of a tweet
- Like social event “suggests”, or constrains, EPA profiles for identities, so too does text in a tweet
- Formalize using quadratic constraints, like ACT does for event model
Strategy for mining affect

For each identity of interest:
- Identify any social events it is involved in
- man (young) -> killed_by -> police_officer
- Find any “sentiment words” (in our sentiment dictionary) in the tweet
- Construct $q$ by summing constraints –
- “Terrible” constraint on police officer ($\phi_{po}$):
  $$(\phi_{po,e} - ter_e)^2 + (\phi_{po,p} - ter_p)^2 + (\phi_{po,a} - ter_a)^2$$

Validation – Semantic model

Fill in the blank (on left out data):
___ rule, boys drool

Baselines:
- **Simple**: Just based on frequency of each identity
- **User**: Laplace-smoothed language model

Metric: Perplexity of identities in left-out data (lower is better)
Validation – Affective model

<table>
<thead>
<tr>
<th>Affective Model</th>
<th>Avg. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>134.744</td>
</tr>
<tr>
<td>User Baseline</td>
<td>127.272</td>
</tr>
<tr>
<td>Our Model</td>
<td>126.042</td>
</tr>
</tbody>
</table>

Fill in the blank (on left out data):

___ rule, boys drool

• Baselines:
  – **Simple**: Tweet-based average using VADER
  – **User**: Simple back-off tweet-based model using VADER

Metric: average rankings of identities in left-out data (lower is better)

Results for Thug

• Top right – affectively similar & semantically related
• N(-a) word semantically, not affectively related
Existing NLP methods - Thug

E.g. deep learning… what words are related to thug?

![Graph showing cosine similarity with related terms like gangsta, thug, goon, thugs, hood, etc.]

Top 10 Related Terms

Cosine Similarity

(Networks of Identities)
**Approach**

- **Twitter data**
  - 150K Twitter users who sent >5 tweets from within the original Arab Spring dataset
- **News data**
  - Original news data
- **Construct common vocabulary; common data format**
- **Run through Bamman et al. Word2Vec embedding model**
- **Determine list of interesting identities**
  - 280 identities prevalent in both datasets
- **Construct network of similarity between these identities for High/Low stability, News/Twitter (4 networks total)**

**Tweets/News Sentences count by country**
Tweets/News articles by unrest level of country

High/Low Civil Unrest Categorization

<table>
<thead>
<tr>
<th>High Unrest Countries</th>
<th>Low Unrest Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain</td>
<td>Qatar</td>
</tr>
<tr>
<td>Iraq</td>
<td>Kuwait</td>
</tr>
<tr>
<td>Iran</td>
<td>Morocco</td>
</tr>
<tr>
<td>Libya</td>
<td>Jordan</td>
</tr>
<tr>
<td>Algeria</td>
<td>Saudi Arabia</td>
</tr>
<tr>
<td>Egypt</td>
<td>Oman</td>
</tr>
<tr>
<td>Syria</td>
<td>United Arab Emirates</td>
</tr>
<tr>
<td>Tunisia</td>
<td>Yemen</td>
</tr>
<tr>
<td>Lebanon</td>
<td></td>
</tr>
</tbody>
</table>
High Unrest, Newspaper data (.7 cutoff, LCC)

High Unrest, Twitter (.65 cutoff, LCC)
Low Unrest, News (.73, LCC)

Low Unrest, Twitter (.65 cutoff)
Conclusion

• Extracting affective stereotypes using “social event networks”
• Extracting a network model of stereotypes
• Networks of identities

• Many different ways to think about identities, text and networks!