Implicit Attitudes, NLP, and the "Real World"

Rob Voigt
Stanford Linguistics

robvoigt@stanford.edu
Other Work

- Quantifying “modernity” in Chinese poetry
  NAACL Comp Ling for Literature 2013
- Discourse-level effects on reference
  who is “you” in reviews?
  ACL 2015
- Gender and pitch in bilinguals
  INTERSPEECH 2016
- Sociophonetic embodiment:
  Body movement and head positioning
  Journal of Sociolinguistics 2016
Today - dissertation time!

Implicit Attitudes

“introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects”

Greenwald and Banaji 1995

NLP allows us to analyze linguistic “actions” at a large scale

...while accounting for crucial aspects of the “real world” social context
Project 1

Racial Disparities in Police Officer Respect

with Nick Camp, Camilla Griffiths, Will Hamilton, David Jurgens, Vinod Prabhakaran, Rebecca Hetey, Dan Jurafsky, and Jennifer Eberhardt
Our Question

Do officers treat White community members with a greater degree of respect than they afford to Blacks?
Police-Community Interaction

- Media focus on explosive incidents
- Research focus on outcomes

but:
- one quarter of adults have contact with the police during the course of a year
  - majority occurring in traffic stops
Respect is Important

● A person who is treated with respect
  ○ ... has more trust in the individual officer’s fairness  
    Tyler and Ho 2001
  ○ ... and the procedural fairness of the institution  
    Tyler and Sunshine 2003
  ○ ... and is more willing to support or cooperate with the police  
    Tyler 1990, Mazerolle et al 2013
Previous work on procedural fairness

- Relies on:
  - citizens’ recollection of past interactions  
    Epp et al 2014
  - researcher observation of officer behavior  

- These are invaluable but indirect
  - ... and presence of researcher may influence police behavior  
    Mastrofski and Parks 1990
Police body camera footage

- Oakland PD has been wearing body cameras since 2010
- Usually used only as evidence
- ... but, a window into everyday behavior!
Our proposal: Footage as Data

- 981 stops by 245 officers in April 2014
  - Drivers: 682 black, 299 white
  - 183 hours of footage

- Professionally transcribed and diarized

- Resulting data set:
  - 36,738 officer utterances, 350k+ words
OFFICER [to dispatch]: Unknown occupant and it's going to be for registration. It should be code four.
OFFICER: Hi.
FEMALE: Hi.
OFFICER: I pulled you over because your registration is expired by almost a year.
FEMALE: Okay, I have the paperwork for it, a moving permit?
OFFICER: I'm sorry?
FEMALE: I have the paperwork for it.
OFFICER: Okay.
Study 1

Perceptions of Officer Treatment from Language
Study 1: Goals

- Can human raters judge respect from officers’ language?
- Are there differences in officer respect towards Black versus White community members?
“Thin Slice” Utterance Rating Task

- Participants (N=70) blind to race labeled 414 officer utterances
  - 10 coders per utterance
  - 4-point Likert scales
    - Respectful, Polite, Friendly, Formal, Impartial
      (high rater agreement $\alpha$s=.73-.91)
Utterance Rating Task

Read the following interaction with a police officer:
The citizen just said:

It's in my glove compartment.

And then the officer says:
Let me take a look at it. How about insurance?

How **impolite** or **polite** was the officer?

- [ ] Very Impolite
- [ ] Somewhat Impolite
- [ ] Somewhat Polite
- [ ] Very Polite
Utterance Rating Task
The Latent Space of Respect

Two PCs explain 93% of the variance:

<table>
<thead>
<tr>
<th></th>
<th>Respect</th>
<th>Formality</th>
</tr>
</thead>
<tbody>
<tr>
<td>variance explained:</td>
<td>71%</td>
<td>22%</td>
</tr>
<tr>
<td>Formal</td>
<td>0.27</td>
<td>0.91</td>
</tr>
<tr>
<td>Friendly</td>
<td>0.47</td>
<td>-0.39</td>
</tr>
<tr>
<td>Polite</td>
<td>0.49</td>
<td>-0.04</td>
</tr>
<tr>
<td>Respectful</td>
<td>0.47</td>
<td>0.03</td>
</tr>
<tr>
<td>Impartial</td>
<td>0.50</td>
<td>-0.11</td>
</tr>
</tbody>
</table>
The Latent Space of Respect

- Race on these dimensions:
Study 2

Modeling Respect with Computational Linguistics
Study 2: Goals

- Develop a computational linguistic model capable of estimating Respect
- Use the human labeled data as supervised training data to learn weights on interpretable features
Methodology

● **Hand-engineered features**
  ○ Lexicons, gazetteers, regexes, dependencies, joint pattern matching ("bald commands")
  ○ Drawn primarily from linguistic and computational work on politeness
    
    Prabhakaran et al 2012, Danescu-Niculescu-Mizil 2013, Krishnan and Eisenstein 2014

● **Statistical Model: simple linear regression**
  ○ log-transformed counts of features per utterance
Feature Weights
Feature Weights

Respect Model Coefficients

Questions
Linguistic Negation
Negative Words
Ask for Agency
Disfluency
Informal Titles
First Names
Hands on the Wheel

Log Odds Ratio by Race

More common in...
Black Stops
White Stops

↑ is p < 0.1, * is p < 0.05, ** is p < 0.01, *** is p < 0.001
<table>
<thead>
<tr>
<th>First Name</th>
<th>Ask For Agency</th>
<th>Questions</th>
<th>Respect Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>[name], can I see that driver's license again?</td>
<td>it- it's showing suspended. Is that- that's you?</td>
<td></td>
<td>-1.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disfluency</th>
<th>Negative Word</th>
<th>Disfluency</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Informal Title</th>
<th>Ask For Agency</th>
<th>Adverbial &quot;Just&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick.</td>
<td></td>
<td>-0.51</td>
</tr>
</tbody>
</table>

"Hands On The Wheel"
<table>
<thead>
<tr>
<th>Example</th>
<th>Respect Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Apology</strong></td>
<td><strong>Introduction</strong></td>
</tr>
<tr>
<td>Sorry to stop you. My name’s Officer [name] with the Police Department.</td>
<td></td>
</tr>
<tr>
<td><strong>Formal Title</strong></td>
<td><strong>Safety</strong></td>
</tr>
<tr>
<td>There you go, ma’am. Drive safe, please.</td>
<td></td>
</tr>
<tr>
<td><strong>Adverbial “Just”</strong></td>
<td><strong>Filled Pause</strong></td>
</tr>
<tr>
<td>It just says that, uh, you’ve fixed it. No problem. Thank you very much, sir.</td>
<td></td>
</tr>
<tr>
<td><strong>Gratitude</strong></td>
<td><strong>Formal Title</strong></td>
</tr>
</tbody>
</table>
Results

- *Respect* model is able to perform roughly like an average annotator

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Adjusted $R^2$</td>
<td>0.258</td>
</tr>
<tr>
<td>Model RMSE</td>
<td>0.840</td>
</tr>
<tr>
<td>Average annotator RMSE</td>
<td>0.842 (range from 0.497 - 1.677)</td>
</tr>
</tbody>
</table>

- *Formality* model is worse but still reasonable

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Adjusted $R^2$</td>
<td>0.190</td>
</tr>
<tr>
<td>Model RMSE</td>
<td>0.882</td>
</tr>
<tr>
<td>Average annotator RMSE</td>
<td>0.764 (range from 0.517 - 1.703)</td>
</tr>
</tbody>
</table>
Study 3

Racial Disparity Across the Entire Dataset
Study 3: Goals

- Do the results from Study 1 hold across an entire month of traffic stops?
- ... even controlling for contextual factors?
## Study 3: Results

<table>
<thead>
<tr>
<th></th>
<th>Respect</th>
<th></th>
<th>Formality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>CI</td>
<td>$\beta$</td>
<td>CI</td>
</tr>
<tr>
<td>Arrest Occurred</td>
<td>-0.00</td>
<td>-0.03 – 0.03</td>
<td>.933</td>
<td>0.01</td>
</tr>
<tr>
<td>Citation Issued</td>
<td>0.04</td>
<td>0.02 – 0.06</td>
<td>&lt;.001</td>
<td>0.01</td>
</tr>
<tr>
<td>Search Conducted</td>
<td>-0.08</td>
<td>-0.11 – 0.05</td>
<td>&lt;.001</td>
<td>-0.00</td>
</tr>
<tr>
<td>Age</td>
<td>0.07</td>
<td>0.05 – 0.09</td>
<td>&lt;.001</td>
<td>0.05</td>
</tr>
<tr>
<td>Gender (F)</td>
<td>0.02</td>
<td>-0.00 – 0.04</td>
<td>.062</td>
<td>0.02</td>
</tr>
<tr>
<td>Race (W)</td>
<td>0.05</td>
<td>0.03 – 0.08</td>
<td>&lt;.001</td>
<td>-0.01</td>
</tr>
<tr>
<td>Officer Race (B)</td>
<td>0.00</td>
<td>-0.03 – 0.04</td>
<td>.884</td>
<td>0.00</td>
</tr>
<tr>
<td>Officer Race (O)</td>
<td>-0.00</td>
<td>-0.04 – 0.03</td>
<td>.809</td>
<td>-0.00</td>
</tr>
<tr>
<td>Officer Race (B) : Race (W)</td>
<td>-0.01</td>
<td>-0.03 – 0.02</td>
<td>.583</td>
<td>0.01</td>
</tr>
<tr>
<td>Officer Race (O) : Race (W)</td>
<td>-0.01</td>
<td>-0.03 – 0.02</td>
<td>.486</td>
<td>-0.00</td>
</tr>
</tbody>
</table>
Interpretation

White community members are 57% more likely to hear an officer say one of the top 10% most respectful utterances in our dataset.

Black community members are 61% more likely to hear an officer say one of the top 10% least respectful utterances in our dataset.
Controls

- Holds even considering:
  - Only “everyday” interactions (no arrest, no search)
  - Crime rate in the area
  - Density of businesses in the area
  - Whether driver race was known before the stop
  - Officer years of experience
Controls - Severity

- We asked OPD officers to rate the stops for severity
  - 1 - very minor (expired registration)
  - 4 - very severe (speeding)

- Black drivers are stopped for less severe offenses
- ... but no impact on respect
Controls - Officer Race

- Surprisingly, not a factor!
Across the Interaction

- Respect rises throughout the interaction
- ... but rises faster for whites
Across the Interaction

- No race effect for Formality
- Officers less formal over the interaction
Conclusions from the first paper

- Confirms community reports: interactions with black community members are more fraught
- Provides concrete strategies for officers
- Cooperation with Oakland to integrate results into procedural justice training
  - ... and we can measure impact
Moving Forward

● Tone of Voice:
  ○ Preliminary results suggest a similar trend

● Community member language:
  ○ Escalation
  ○ Compliance, politeness

● Other Departments