Racial Disparities in Police Language Use

Rob Voigt

CS124

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 Black community members?
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 & Ho, 2001)
  ○ ... and the procedural fairness of the institution  
    (Tyler & 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 (Mastrofski et al, 2009; Dai, Frank, & Sun, 2011; Jonathan-Zamir, Mastrofski, & Moyal, 2015; Mastrofski et al, 2016)

● These are invaluable but indirect
  ○ … and presence of researcher may influence police behavior (Mastrofski & Parks, 1990)
This work: 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

- We look at all traffic stops of black and white community members in April 2014
- 981 stops by 245 officers
  - Drivers: 682 black, 299 white
  - 183 hours of footage
Transcribed dataset

- Professional transcribers
  - Underwent background checks by OPD
  - Watched videos
  - Transcribed words
  - Diarized (who is talking to whom)

- Resulting data set:
  - 36,738 officer utterances, 350k+ words
0:00:00 0:00:09 OFFICER [to dispatch]: Unknown occupant and it's going to be for registration. It should be code four.
0:00:20 0:00:20 OFFICER: Hi.
0:00:20 0:00:20 FEMALE: Hi.
0:00:21 0:00:23 OFFICER: I pulled you over because your registration is expired by almost a year.
0:00:25 0:00:28 FEMALE: Okay, I have the paperwork for it, a moving permit?
0:00:28 0:00:28 OFFICER: I'm sorry?
0:00:29 0:00:30 FEMALE: I have the paperwork for it.
0:00:30 0:00:31 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?
Utterance Rating Task

- Participant Coders (N=70) blind to citizen race labeled 414 unique officer utterances
  - 10 coders per utterance
  - 4-point Likert scales (high rater agreement $\alpha$s=.73-.91)

  Respectful, Polite, Friendly, Formal, and Impartial
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

Community Member Race

- black
- white
The Latent Space of Respect

- The questions we asked are all very correlated

- How do we discover the underlying construct?

- Principal Component Analysis
  - Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components
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:
Next Steps: Scaling up

● We ran this whole experiment and only labeled 414 officer utterances!
  ○ What about the remaining 35k+?
  ○ … and that’s only 1 month, and only traffic stops

● Could we automate this process to be able to look at large amounts of text data?
  ○ … and control for many contextual variables
Study 2

Modeling Respect with Computational Linguistics
Study 2: Goals

- Use linguistic theories of respect and social distance to develop linguistic features that can be detected in transcripts
- Use the human labeled data as supervised training data to learn weights on these interpretable features
Linguistic Models of Respect


- “Positive and Negative Politeness”

- Mitigating face-threatening acts
  - Positive face
    - hearer’s self-image
  - Negative face
    - hearer’s freedom of action
Linguistic Models of Respect

- Linguistic theories of politeness focus on **requests**
  - Ordering you to do something is face-threatening

- **Negative Politeness:**
  - Minimize (maximize) my request
  - put on record that it’s an imposition on you (vs. ignore the impact on you)

- **Positive Politeness:**
  - Emphasize your value (vs. deemphasize your value)
  - Emphasize (vs. deemphasize) my good relationship with you
Features for Negative Politeness

- **Apologizing**
  - “sorry”, “oops”, “my fault”, “excuse me”

- **Gratitude**
  - “thank.*”, “appreciate”

- **Imposition minimizers**
  - “it’s ok”, “don’t worry”, “no big deal”, “you’re good”

- **Hedges (LIWC “Tentative words”)**
  - “maybe”, “a little”, “kind of”, “sort of”

- **Give Agency**
  - “let you”, “allow you”, “you may”, “you can”

- **Negative impoliteness: Control their actions**
  - “hands on the wheel”
Features for Positive Politeness

- **Formal versus informal titles**
  - “sir”, “ma’am”, “Mr.”... versus “bro”, “man”, “dude”

- **Last names, first names**
  - [Frequent names from US Census]

- **Introductions**
  - “Hello”, “My name is”, “I’m Officer X”

- **Mentioning safety**
  - “safe.*”

- **Positive words**
  - “good”, “great”, “awesome”, “perfect”
Regexes!

Apologies
“( sorry | oops | woops | excuse me | forgive me | apologies | my bad | apologize | my fault )”

Time Minimizing
“ (a|one|a few) (minute|min|second|sec|moment)s? | this[^,.?!]+quick | right back ”
Methodology

- Hand-engineered features
  - Lexicons
  - Regular expressions
  - Dependency-based rules
  - More complex functions (“bald commands” etc)

- Statistical Model: simple linear regression
  - log-tranformed counts of features per utterance
  - stepwise removal of uninformative features
Results

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</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>
Feature Weights

Respect Model Coefficients

- Apologizing
- For You
- Gratitude
- Reassurance
- Last Names
- Formal Titles
- For Me
- Safety
- Give Agency
- Filled Pauses (Um/Uh)
- Adverbial Just
- Positive Words
- Hedges
- Introductions

 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
Feature Weights

Respect Model Coefficients

Perceived as more...
Disrespectful  Respectful

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><strong>Example</strong></th>
<th><strong>Respect Score</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Name</strong></td>
<td><strong>Ask For Agency</strong></td>
</tr>
<tr>
<td>[name], can I see that driver's license again?</td>
<td>It- it's showing suspended. Is that- that's you?</td>
</tr>
<tr>
<td><strong>Disfluency</strong></td>
<td><strong>Negative Word</strong></td>
</tr>
<tr>
<td><strong>Informal Title</strong></td>
<td><strong>Ask For Agency</strong></td>
</tr>
<tr>
<td>All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick.</td>
<td>-0.51</td>
</tr>
<tr>
<td>&quot;HANDS ON THE WHEEL&quot;</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td>Respect Score</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
</tr>
<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>
</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></th>
<th>Formality</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\beta)</td>
<td>CI</td>
<td>(p)</td>
<td>(\beta)</td>
<td>CI</td>
<td>(p)</td>
</tr>
<tr>
<td>Arrest Occurred</td>
<td>-0.00</td>
<td>-0.03 - 0.03</td>
<td>.933</td>
<td>0.01</td>
<td>-0.02 - 0.04</td>
<td>.528</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>
<td>-0.01 - 0.03</td>
<td>.209</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>
<td>-0.03 - 0.02</td>
<td>.848</td>
</tr>
<tr>
<td>Age</td>
<td>0.07</td>
<td>0.05 - 0.09</td>
<td>&lt;.001</td>
<td>0.05</td>
<td>0.03 - 0.07</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Gender (F)</td>
<td>0.02</td>
<td>-0.00 - 0.04</td>
<td>.062</td>
<td>0.02</td>
<td>0.00 - 0.04</td>
<td>.025</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>
<td>-0.04 - 0.01</td>
<td>.236</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>
<td>-0.03 - 0.03</td>
<td>.987</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>
<td>-0.03 - 0.02</td>
<td>.783</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>
<td>-0.01 - 0.03</td>
<td>.188</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>
<td>-0.02 - 0.02</td>
<td>.928</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 officers to rate the stops for their severity
  - 1 - very minor (expired registration)
  - 4 - very severe (speeding)

- Black drivers are stopped for less severe offenses

- ... but no impact on respect
Surprisingly, not a factor! (including homophily)
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
Implications for Escalation

- Looked at LIWC Anger/Swear words used by Community Member

- When officers are more respectful, are drivers less angry?
Moving Forward

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

● **Community member language:**
  ○ Escalation
  ○ Compliance, politeness

● **Other Departments**
Conclusion

- NLP can help us understand abstract concepts like “Respect” and quantify them
- The techniques you learned in this class are powerful and applicable to the real world!

nlp.stanford.edu/robvoigt

Thank you!

Questions welcome!
robvoigt@stanford.edu