Tweetment Effects on the Tweeted: An Experiment to Reduce Twitter Harassment

Kevin Munger

NYU

December 1, 2016
Prepared for Yale Human Nature Lab
Project Outline

- Find harassers on Twitter
Project Outline

- Find harassers on Twitter
- Randomize assignment 2x2
  - In-group v Out-group (race of bot)
  - High Status v Low Status (number of followers)
Project Outline

- Find harassers on Twitter
- Randomize assignment 2x2
  - In-group v Out-group (race of bot)
  - High Status v Low Status (number of followers)
- Offensiveness as outcome variable
Project Outline

- Find harassers on Twitter
- Randomize assignment 2x2
  - In-group v Out-group (race of bot)
  - High Status v Low Status (number of followers)
- Offensiveness as outcome variable
- See how long the treatment persists
Why does it happen?

- Prejudice
Why does it happen?

- Prejudice
- Social Identity model of De-Individuation (SIDE)
Why does it happen?

- Prejudice
- Social Identity model of De-Individuation (SIDE)
- Girard’s theory of mimetic desire
State of the art

- Experiments in the lab
  - Convenience samples
  - Short time frame
  - In the lab
State-of-the-art My Approach

- Experiments in the lab vs. Experiment in the “field”
  - Convenience samples: Sample of real, consistent harassers
  - Short time frame: Continuous and unbounded time frame
  - In the lab: In the same context as the harassment
Find harassers

- Needs to be fast, and accurate

[Chen et al, 2012]
Find harassers

- Needs to be fast, and accurate
- Don’t care about recall

In the presence of strongly offensive language, a dictionary of slurs is best (Chen et al, 2012)
Find harassers

- Needs to be fast, and accurate
- Don’t care about recall
- In the presence of strongly offensive language, a dictionary of slurs is best (Chen et al, 2012)
Detection “Algorithm”

Include

StreamR finds a tweet with “nigger”

Is the tweet with the slur an “@”-reply?

Is the user’s offensiveness score above the 90th percentile?

Is the user an adult white man?

Manually inspect history of interaction between user and subject of harassment—do they appear to be genuinely harassed?

Exclude

EXCLUDE

EXCLUDE

EXCLUDE

EXCLUDE

Assign to a treatment condition subject to balance constraints
Apply Treatment

13 Sep 2015

@... don't be a nigger

---

@... Hey man, just remember that there are real people who are hurt when you harass them with that kind of language
Treatment uptake
Hypotheses

All hypotheses have been pre-registered through EGAP.

Hypothesis

The ranking of the magnitudes of the decrease in harassment will be:

\[
\text{In-group/High status} > \text{In-group/Low status} > \text{Out-group/High status} > \text{Out-group/Low status}.
\]
All hypotheses have been pre-registered through EGAP.

**Hypothesis**

The ranking of the magnitudes of the decrease in harassment will be:  
\[ \text{In-group/High status} > \text{In-group/Low status} > \text{Out-group/High status} > \text{Out-group/Low status}. \]

**Table:** Experimental Design and Hypothesized Effect Sizes

<table>
<thead>
<tr>
<th></th>
<th>In-group</th>
<th>Out-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low followers</td>
<td>Medium effect</td>
<td>Small effect</td>
</tr>
<tr>
<td>High followers</td>
<td>Large effect</td>
<td>Medium effect</td>
</tr>
</tbody>
</table>
Hypotheses

All hypotheses have been pre-registered through EGAP.

Hypothesis

The ranking of the magnitudes of the decrease in harassment will be:
In-group/High status > In-group/Low status > Out-group/Low status > Out-group/High status.

Table: Experimental Design and Hypothesized Effect Sizes

<table>
<thead>
<tr>
<th></th>
<th>In-group</th>
<th>Out-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low followers</td>
<td>Medium effect</td>
<td>Small effect</td>
</tr>
<tr>
<td>High followers</td>
<td>Large effect</td>
<td>Medium effect</td>
</tr>
</tbody>
</table>
Hypotheses

All hypotheses have been pre-registered through EGAP.

Hypothesis

The ranking of the magnitudes of the decrease in harassment will be:

In-group/High status \( > \) In-group/Low status \( > \) Out-group/Low status.

Table: Experimental Design and Hypothesized Effect Sizes

<table>
<thead>
<tr>
<th></th>
<th>In-group</th>
<th>Out-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low followers</td>
<td>Medium effect</td>
<td>Small effect</td>
</tr>
<tr>
<td>High followers</td>
<td>Large effect</td>
<td>Medium effect</td>
</tr>
</tbody>
</table>

Hypothesis

The decrease in offensive language will be smaller in subjects who provide less information in their profile.
Change in Racist Language: Full Sample (242)

1 Week

2 Weeks

1 Month

Change in Average Daily Slur Use

Log Followers

Out-Group/High

In-Group/High

Out-Group/Low

In-Group/Low

-0.6 -0.4 -0.2 0.0 0.2 0.4

Change in Average Daily Slur Use

-0.6 -0.4 -0.2 0.0 0.2 0.4

-0.6 -0.4 -0.2 0.0 0.2 0.4
Change in Racist Language: Non-Anonymous Sample (84)

1 Week

2 Weeks

1 Month

Change in Average Daily Slur Use

Log Followers

Out-Group/High

In-Group/High

Out-Group/Low

In-Group/Low

−0.6 −0.4 −0.2 0.0 0.2 0.4

−0.6 −0.4 −0.2 0.0 0.2 0.4

−0.6 −0.4 −0.2 0.0 0.2 0.4

Change in Average Daily Slur Use
Science Moves Quickly
In Real World Terms

My intervention caused the 50 subjects in the most effective condition to tweet the word “nigger” an estimated 186 fewer times in the month after treatment.
Ongoing Project

- Political civility
Ongoing Project

- Political civility
- Incivility demobilizes and polarizes
A Visual Overview

Donald J. Trump @realDonaldTrump · Oct 11
In Texas now, leaving soon for BIG rally in Florida!

NSA agent jim @realDonaldTrump HOPE U DIE

Sean @realDonaldTrump typical libtard go fuck yourself
1:25 AM · 12 Oct 2016

Matthew @Matthew FEELINGS
You shouldn't use language like that. Republicans need to remember that our opponents are real people, with real feelings.
Detection

Yes

StreamR finds a tweet with “@realDonaldTrump” or “@HillaryClinton”

Is the tweet an “@”-reply to someone besides Trump or Clinton?

Calculate aggression score; is tweet in top 10% most aggressive?

Does the potential subject appear to be an adult speaking English, with a Twitter account at least 2 months old?

Is the incivility directed at someone besides a journalist or other political actor?

Is the incivility directed at someone who expressed a different political viewpoint?

Assign to a treatment condition subject to balance constraints

No

EXCLUDE

EXCLUDE

EXCLUDE

EXCLUDE

EXCLUDE
Manipulations

- All bots are white men with many followers
Manipulations

- All bots are white men with many followers
- Trump-critical subjects get a tweet from a Democrat or Hillary bot
Manipulations

- All bots are white men with many followers
- Trump-critical subjects get a tweet from a Democrat or Hillary bot
- Test effectiveness of three types of messages
Manipulations

- All bots are white men with many followers
- Trump-critical subjects get a tweet from a Democrat or Hillary bot
- Test effectiveness of three types of messages
- Ideologically scale subjects (Barberá, 2015) and look for heterogeneous effects
Tweetment Variations

- Different rhetoric to appeal to different moral frameworks
Tweetment Variations

- Different rhetoric to appeal to different moral frameworks
  - Authority moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility."
  - Care moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings."
  - Placebo message: "Remember that everything you post here is public. Everyone can see that you tweeted this."
Different rhetoric to appeal to different moral frameworks

- Authority moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility."

- Care moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings."
Tweetment Variations

Different rhetoric to appeal to different moral frameworks

- Authority moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility."

- Care moral foundation: "You shouldn’t use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings."

- Placebo message: "Remember that everything you post here is public. Everyone can see that you tweeted this."
Preliminary Results

Response Rates by Treatment (N=224)

- Feelings
- Rules
- Public
Preliminary Results

Percentage of Conciliatory Response (N=72)

LEFTIST                          RIGHTIST
0.00 0.10 0.20 0.30

Feelings                      Rules  Public  Feelings  Rules  Public

LEFTIST

RIGHTIST
An Optimistic Note

Veteran @name 11h
@realDonaldTrump a nasty woman beat the piss out of you in front of millions, yet u claim victory. You are a bad hombre

1st Sgt. @name 11h
@name @realDonaldTrump repeat that statement when ure head gets sawed off.

Veteran @name 11h
@name @realDonaldTrump u know what's absolutely hilarious, is I've patrolled Mosul and dealt w al queda snipers. I don't fear shit
An Optimistic Note

1st Sgt. [Username] 11h
@ [Username] realDonaldTrump I spilled blood in the desert for two tours
one in storm and one in shield, ure point is.

1st Sgt. [Username] 11h
@ [Username] realDonaldTrump its ure.vote. ure right. I just don't see the
logic. Good luck

Veteran [Username] 11h
@ [Username] realDonaldTrump I'm happy to argue, but not to be called a
disgrace. I'm not on day zero of basic

1st Sgt. [Username] 11h
@ [Username] realDonaldTrump ure opinion is ures and mine is mine. Your
right tho I should of kept the fact of ure disgrace to myself

Veteran [Username] 11h
@ [Username] realDonaldTrump I don't really take it personally if it's another
vet. I'm sorry i was upset
An Optimistic Note

1st Sgt. @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump sorry for the comment

Veteran @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump me too. I have mad respect for you. one tour was enough

1st Sgt. @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump My grandfather, same division. He had it tougher though. Omaha Beach ww2

Veteran @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump jesus, yes.

1st Sgt. @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump These wars are gonna be the longest and most difficult to win. Its like a widespread Vietnam

Veteran @ [redacted] @realDonaldTrump 11h
@ [redacted] @realDonaldTrump absolutely. 15 years and counting, way too much. At least there is some hope in Iraq

1st Sgt. @ [redacted] @realDonaldTrump
@ [redacted] @realDonaldTrump With the right leader and Generals in place
Thanks for your comments, and for listening!

- km2713@nyu.edu
- @kmmunger (no harassment, please)
## Attrition rates

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline # of subjects</td>
<td>40</td>
<td>49</td>
<td>44</td>
<td>50</td>
<td>48</td>
</tr>
<tr>
<td># with &gt; 1 Post-treatment tweets</td>
<td>40</td>
<td>46</td>
<td>42</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td># with &gt; 25 Post-treatment tweets</td>
<td>40</td>
<td>34</td>
<td>33</td>
<td>35</td>
<td>43</td>
</tr>
<tr>
<td>Attrition %, &lt; 25 tweets</td>
<td>10%</td>
<td>18%</td>
<td>16%</td>
<td>18%</td>
<td>4%</td>
</tr>
</tbody>
</table>

The number of subjects who tweeted more than 1 or 25 times after the application of the treatment.