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~~ 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

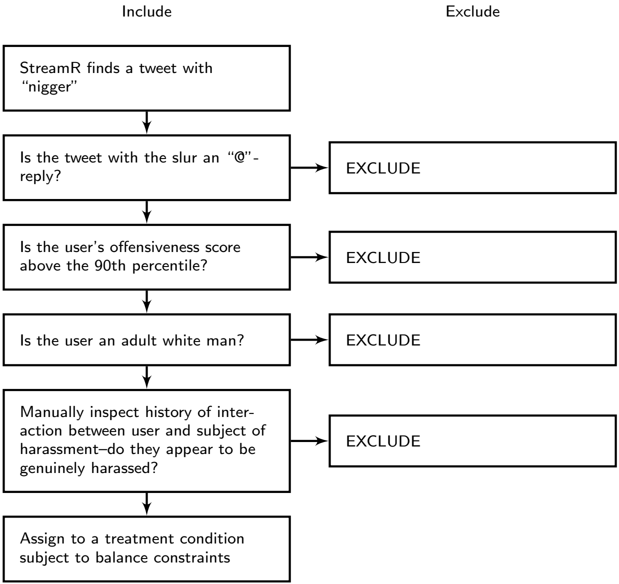
Find harassers

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

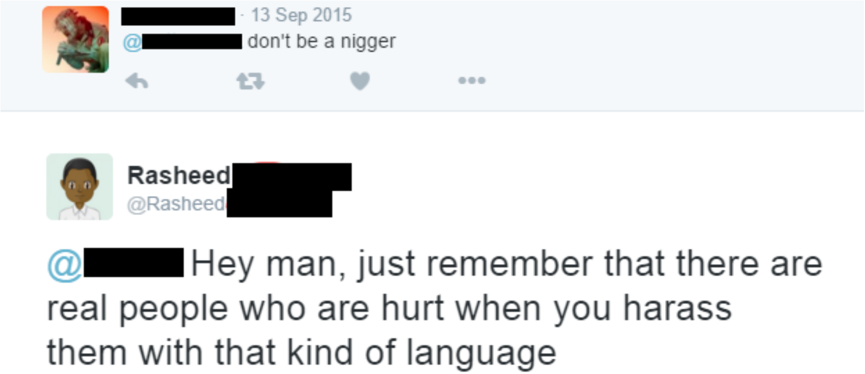
Find harassers

- 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"



Apply Treatment



Treatment uptake

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 $> \frac{In-group/Low\ \ status}{Out-group/High\ \ status} >$ *Out-group/Low status.*

Hypotheses

All hypotheses have been pre-registered through EGAP.

Hypothesis

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

Table: Experimental Design and Hypothesized Effect Sizes

	In-group	Out-group
Low followers	Medium effect	Small effect
High followers	Large effect	Medium effect

Hypotheses

All hypotheses have been pre-registered through EGAP.

Hypothesis

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

Table: Experimental Design and Hypothesized Effect Sizes

	In-group	Out-group
Low followers	~~Medium effect~~	~~Small effect~~
High followers	Large effect	~~Medium effect~~

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 $> \frac{In-group/Low\ \ status}{Out-group/High\ \ status} >$ *Out-group/Low status.*

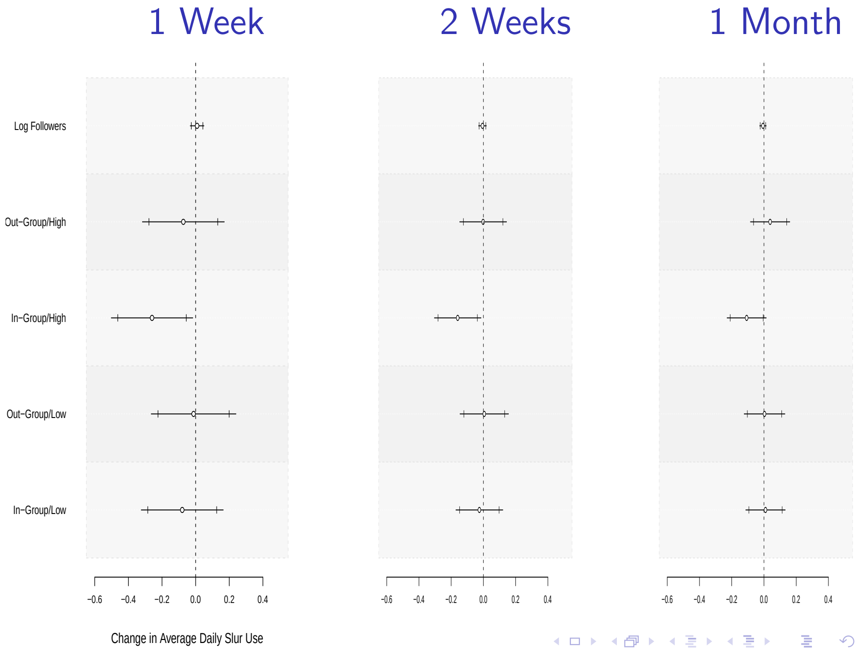
Table: Experimental Design and Hypothesized Effect Sizes

	In-group	Out-group
Low followers	~~Medium effect~~	~~Small effect~~
High followers	Large effect	~~Medium effect~~

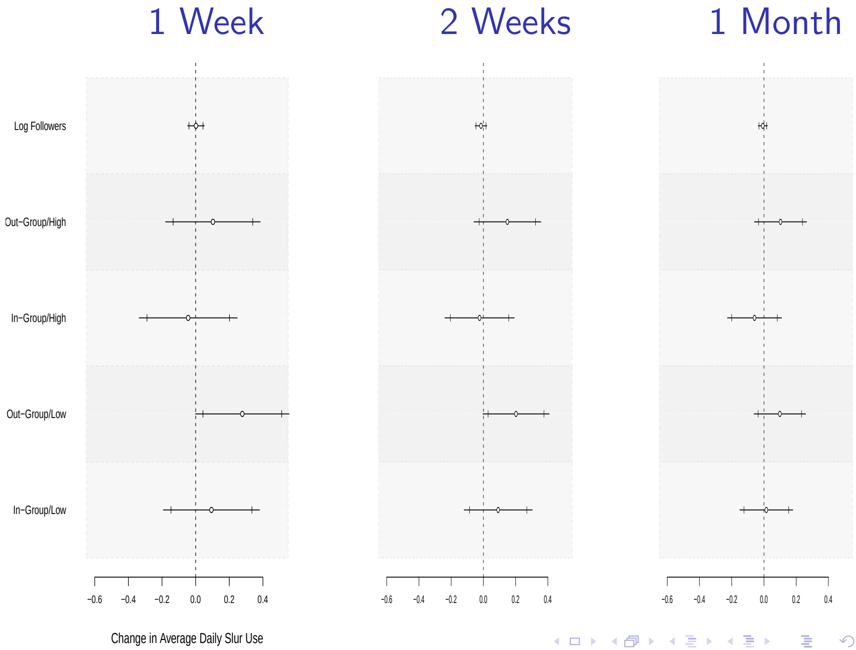
Hypothesis

The decrease in offensive language will be smaller in subjects who provide less information in their profile.

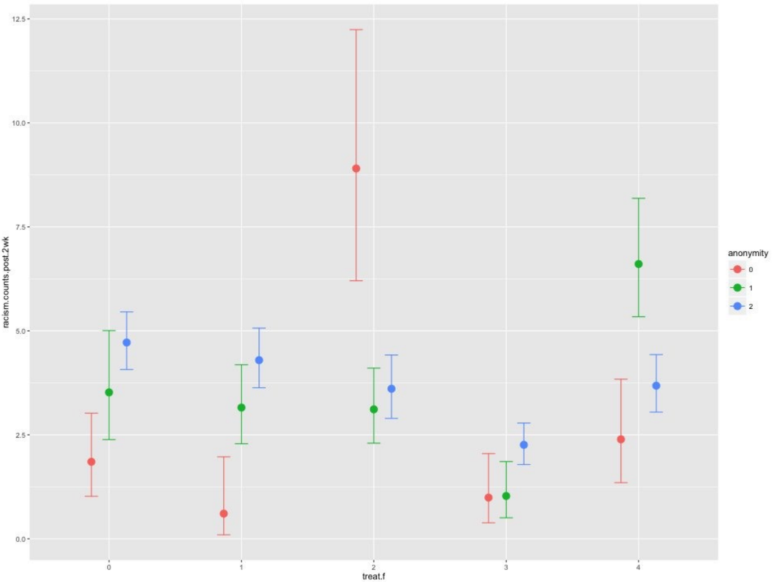
Change in Racist Language: Full Sample (242)



Change in Racist Language: Non-Anonymous Sample (84)



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

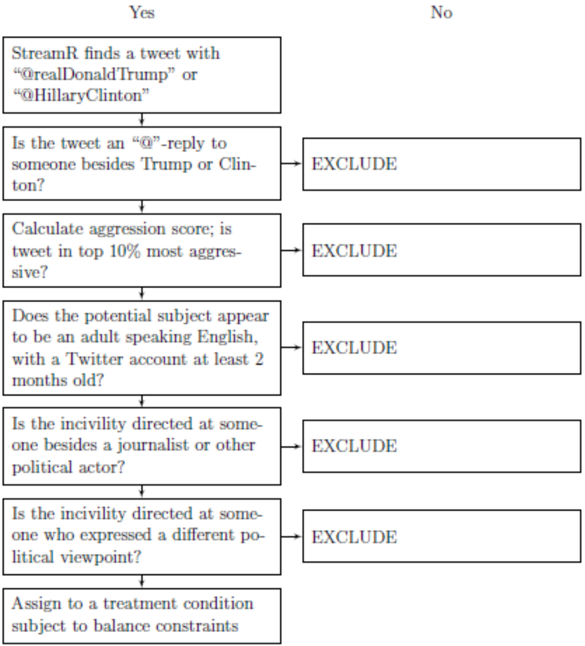
- Political civility

Ongoing Project

- Political civility
- Incivility demobilizes and polarizes

A Visual Overview

Detection


Yes
No
StreamR finds a tweet with "@realDonaldTrump" or "@HillaryClinton"
Is the tweet an "@"-reply to someone besides Trump or Clinton?
EXCLUDE
Calculate aggression score; is tweet in top 10% most aggressive?
EXCLUDE
Does the potential subject appear to be an adult speaking English, with a Twitter account at least 2 months old?
EXCLUDE
Is the incivility directed at someone besides a journalist or other political actor?
EXCLUDE
Is the incivility directed at someone who expressed a different political viewpoint?
EXCLUDE
Assign to a treatment condition subject to balance constraints

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."

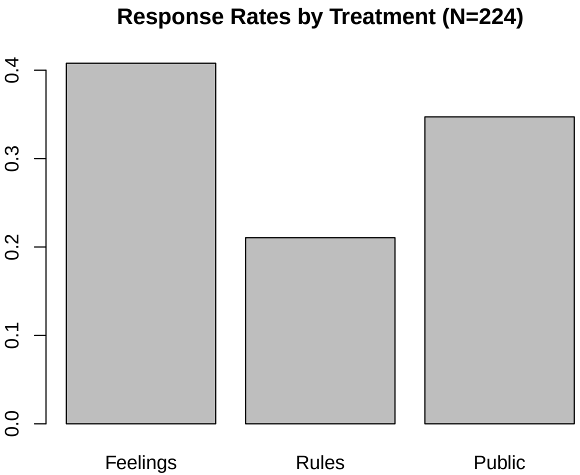
Tweetment Variations

- 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."

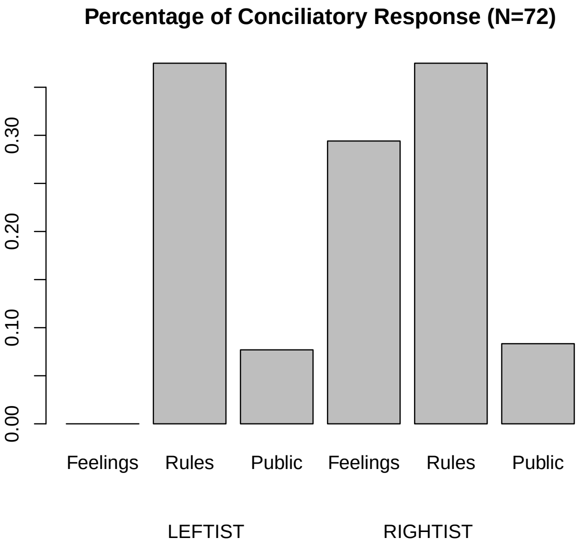
Tweetment Variations

- 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



Preliminary Results



An Optimistic Note

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

62 437

1st Sgt. [redacted] @[redacted] · 11h
@[redacted] @realDonaldTrump repeat that statement when ure head gets sawed off.

3

Veteran [redacted] @[redacted] · 11h
@[redacted] @realDonaldTrump u know what's absolutely hilarious, is I've patrolled Mosul and dealt w al queda snipers. I don't fear shit

37

An Optimistic Note

1st Sgt. [redacted] 11h

@[redacted] @realDonaldTrump I spilled blood in the desert for two tours one in storm and one in shield, ure point is.

1st Sgt. [redacted] 11h

@[redacted] @realDonaldTrump its ure.vote. ure right. I just don't see the logic. Good luck

3

Veteran [redacted] 11h

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

5

1st Sgt. [redacted] 11h

@[redacted] @realDonaldTrump ure opinion is ures and mine is mine. Your right tho I should of kept the fact of ure disgrace to myself

2

Veteran [redacted] 11h

@[redacted] @realDonaldTrump i don't really take it personally if it's another vet. I'm sorry i was upset

5

An Optimistic Note

1st Sgt.[redacted] · 11h
@[redacted] @realDonaldTrump sorry for the comment
2

Veteran [redacted] · 11h
@[redacted] @realDonaldTrump me too. I have mad respect for you, one tour was enough
4

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

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

1st Sgt.[redacted] · 11h
@[redacted] @realDonaldTrump These wars are gonna be the longest and most difficult to win. Its like a widespread Vietnam
2

Veteran [redacted] · 11h
@[redacted] @realDonaldTrump absolutely. 15 years and counting, way too much. At least there is some hope in Iraq
2

1st Sgt.[redacted]
@[redacted]
Follow

@[redacted] @realDonaldTrump With the right leader and Generals in place

LIKE
1

Thanks for your comments, and for listening!

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

Attrition rates

	Control	A	B	C	D
Baseline # of subjects	40	49	44	50	48
# with $>$ 1 Post-treatment tweets	40	46	42	47	47
# with $>$ 25 Post-treatment tweets	40	34	33	35	43
Attrition %, $<$ 25 tweets	10%	18%	16%	18%	4%

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