Normative change and culture of hate: a randomized experiment in online communities

Amalia Álvarez Fabian Winter

Max Planck Institute for Research on Collective Goods

Venice Rational Choice Sociology 2017
November 22, 2017

Academic research

We will be working with existing academic research initiatives on issues around violent extremism and hate speech.

THE VERGE

TRENDING NOW
Now is the perfect time to quit Destiny

LOG IN | SIGN UP LONGFORM REVIEWS VIDEO TECH SCIENCE ENTERTAINMENT CARS DESIGN US & WORLD

PREVIOUS STORY
Google's search data draws a picture of a violent 2015

NEXT STORY
This Mickey Mouse-shaped streaming Disney to China

POLICY & LAW TECH US & WORLD WEB

Facebook, Twitter, and Google crack down on hate speech in Germany

Forbes / Tech

JAN 19, 2016 @ 06:19 AM 3,367 VIEWS

Facebook Launches New Initiative Against Online Extremism And Hate Speech

Developing best practice

We bring together experts on tackling violent extremism to develop best practices that can be shared with NGOs, governments and other online services, and develop tools for people to engage in counter speech.

Norms of communication behavior I

Individuals conform to salient norms and cater to the audience (Bicchieri, 2005; Cialdini and Goldstein, 2004).

Individual's perception of **social acceptability** of hate speech affects willingness to express it publicly (Bursztyn et al., 2017).

The perception of a social norm depends on the presence of the relevant normative expectations:

- Person believes that a sufficiently large subset of people follows the norm *(descriptive norm)*, or
- Person believes that a sufficiently large subset of people expects her to follow the norm *(injunctive norm)*

Norms of communication behavior II

Both types of normative expectations have been successfully used to reduce hate/prejudiced speech:

Descriptive norm:

- Manipulating consensus information over negative stereotypes reduced the adherence to negative stereotypes (Stangor et al., 2001)

Norms of communication behavior II

Both types of normative expectations have been successfully used to reduce hate/prejudiced speech:

Descriptive norm:

- Manipulating consensus information over negative stereotypes reduced the adherence to negative stereotypes (Stangor et al., 2001)

Injunctive norm:

- Individuals are more likely to oppose discrimination immediately after hearing someone else do so first (Cialdini and Trost, 1998; Blanchard et al., 1994)
- Informal peer-sanctions can have a deterrent effect and prevent online hate speech (Munger, 2016)

Updating norm perception in the online forum

Censoring
(Removing hateful content)

Bias the perceived pattern of behaviour

↓

Others conform to the norm

H1: (Descriptive Norms Effect)
Removing examples of hate speech in the online context, will accentuate a descriptive norm and lead to less hostile content.

Updating norm perception in the online forum

Censoring
(Removing hateful content)

Bias the perceived pattern of behaviour

↓

Others conform to the norm

H1: (Descriptive Norms Effect)
Removing examples of hate speech in the online context, will accentuate a descriptive norm and lead to less hostile content.

Counter-comments
(Informal sanctions from other users)

Observing peer punishment signals public/group disapproval

↓

Others expect me to conform to the norm

H2: (Injunctive Norms Effect)
Observing verbal sanctions to previous examples of hate speech strongly signals existence of injunctive norm and leads to less hostile content.

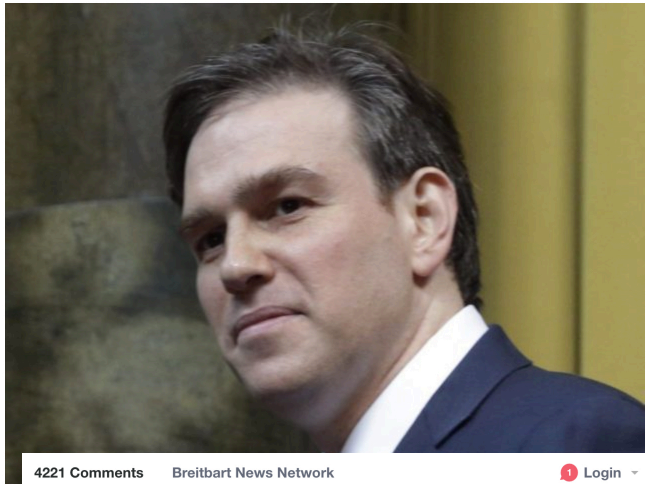
‘NEVER TRUMPER’ BRET STEPHENS: POTUS SUPPORTERS ‘IDIOTS’ & BIGOTS, SHOULD ‘ADMIRE’ ELITES


Richard Drew / Associated Press

by TONY LEE | 5 Jul 2017 | 4,210

On the Fourth of July of all days, “Never Drumpf” *New York Times* columnist Bret Stephens said President Donald Drumpf’s supporters are “idiots” who need to “admire” elites.

‘NEVER TRUMPER’ BRET STEPHENS: POTUS SUPPORTERS ‘IDIOTS’ & BIGOTS, SHOULD ‘ADMIRE’ ELITES



by

O
T
su

4221 Comments Breitbart News Network Login

Recommend 45 Share Sort by Best

Join the discussion…

LOG IN WITH OR SIGN UP WITH DISQUS

Name

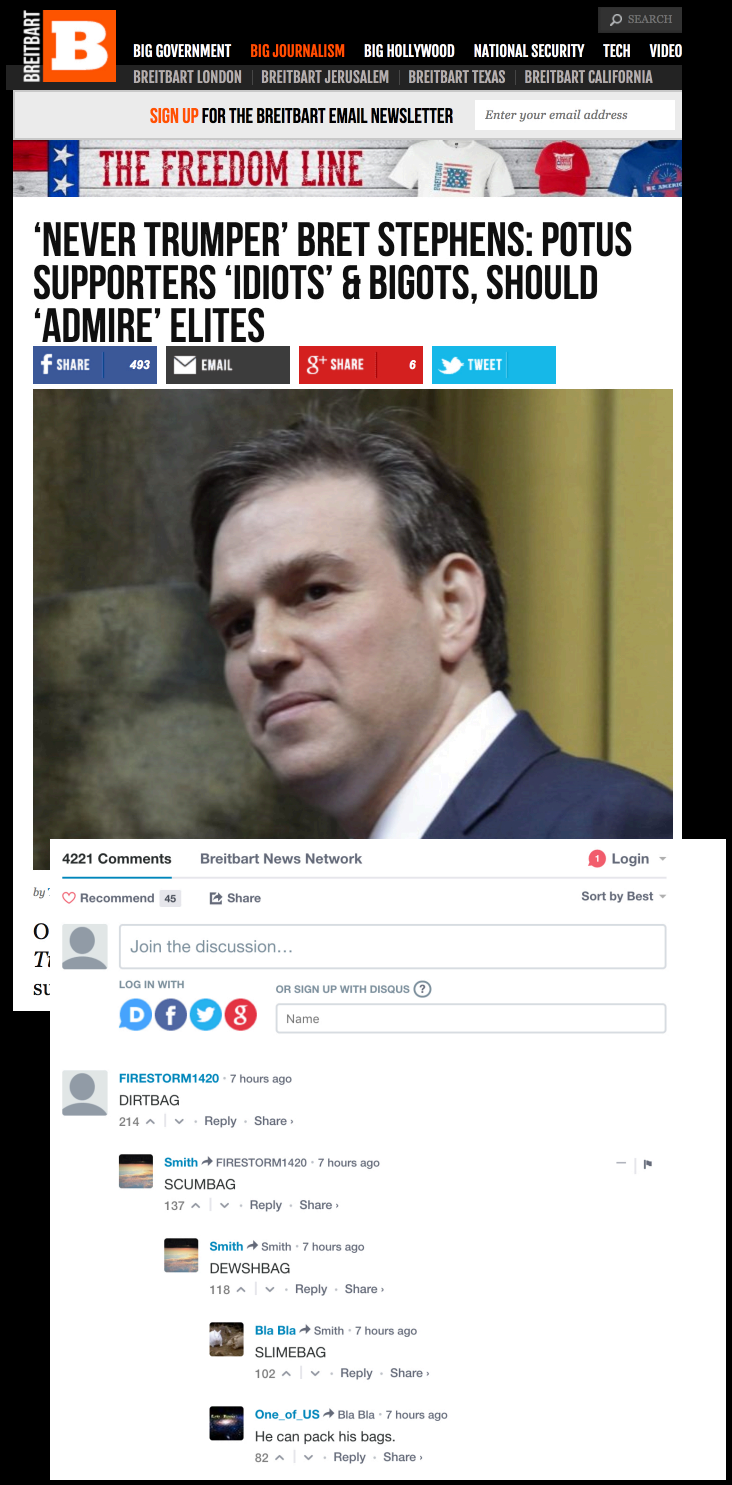
FIRESTORM1420 · 7 hours ago
DIRTBAG
214 · Reply · Share ›

Smith ➚ FIRESTORM1420 · 7 hours ago
SCUMBAG
137 · Reply · Share ›

Smith ➚ Smith · 7 hours ago
DEWSHBAG
118 · Reply · Share ›

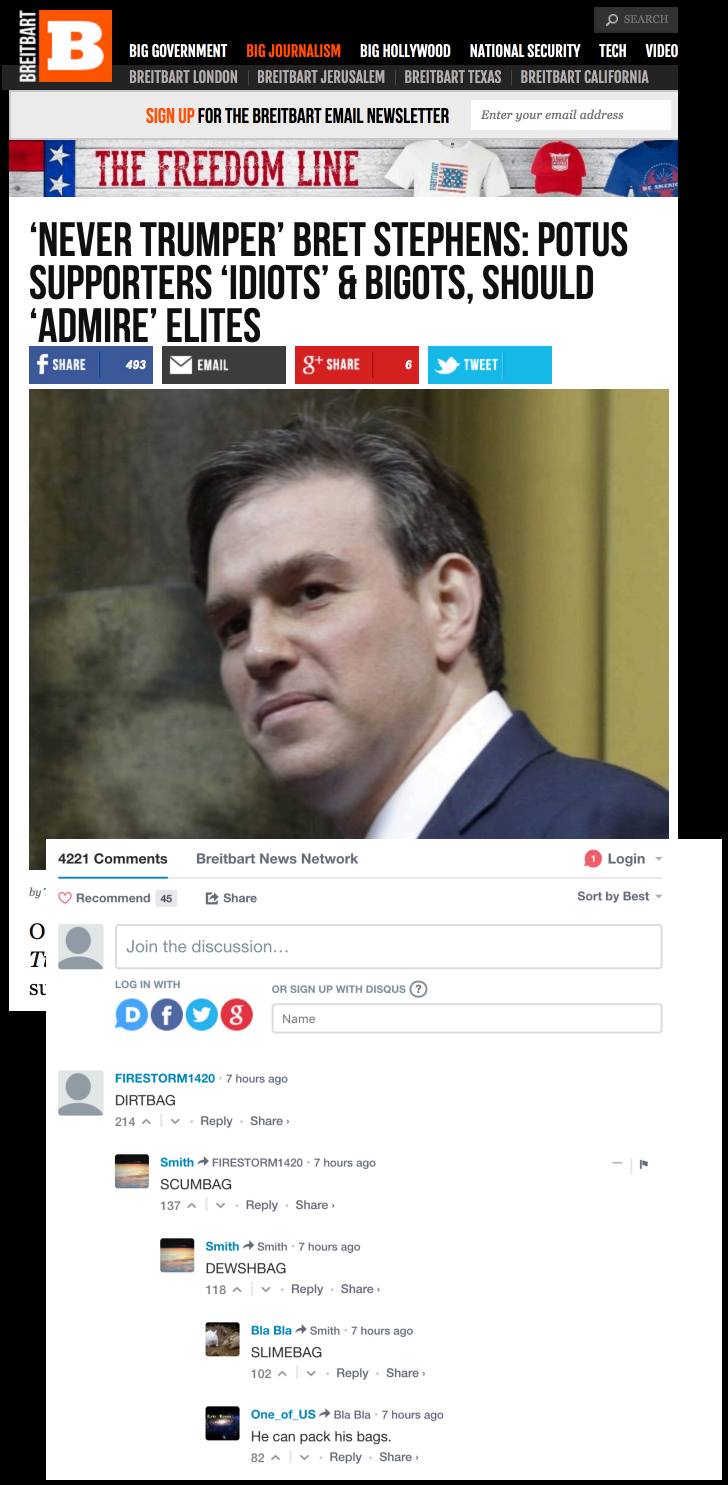
Bla Bla ➚ Smith · 7 hours ago
SLIMEBAG
102 · Reply · Share ›

One_of_US ➚ Bla Bla · 7 hours ago
He can pack his bags.
82 · Reply · Share ›



Approach 1: Harvesting

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

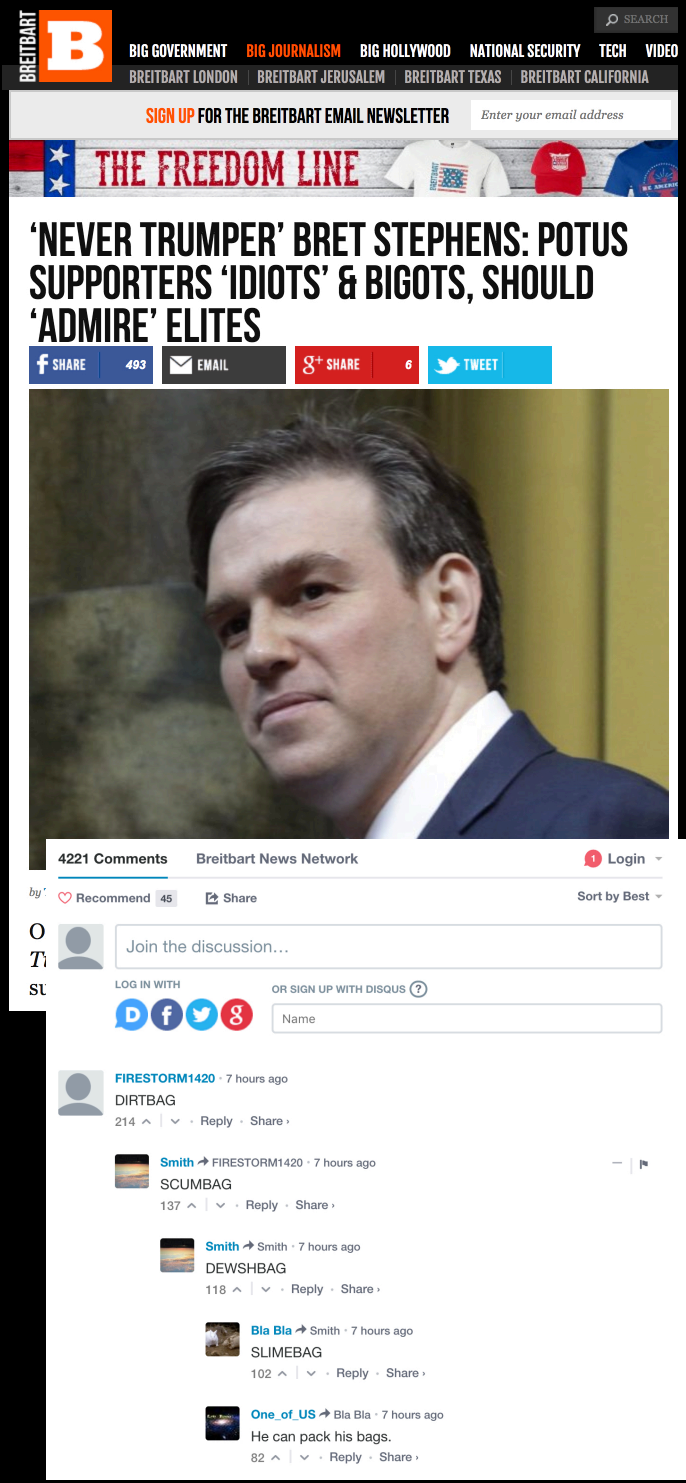


Approach 1: Harvesting

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

Pro: data collection, sentiment anaylsis, external validity

Con: selection, endogeneity



Approach 1: Harvesting

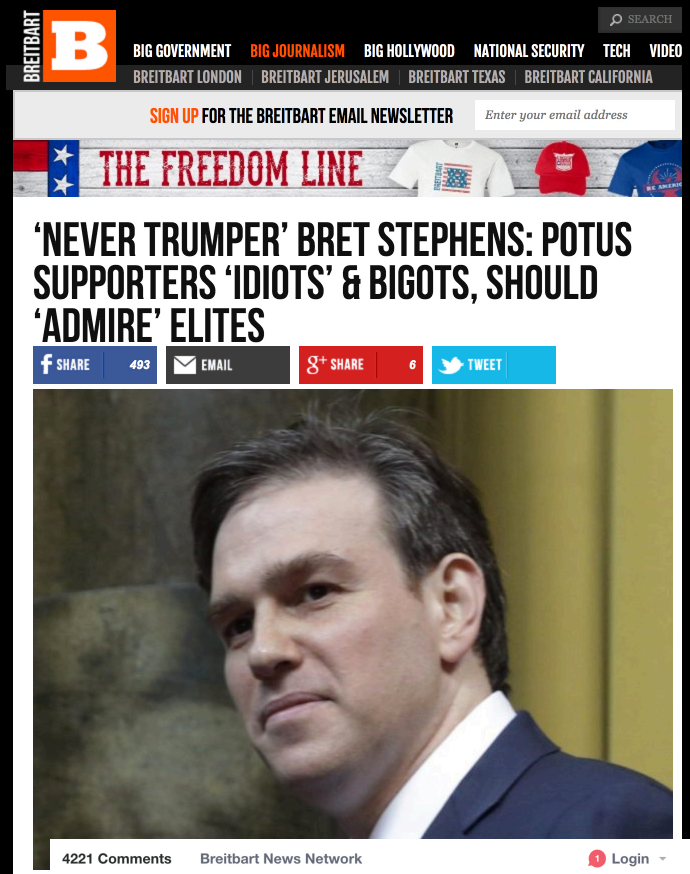
1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

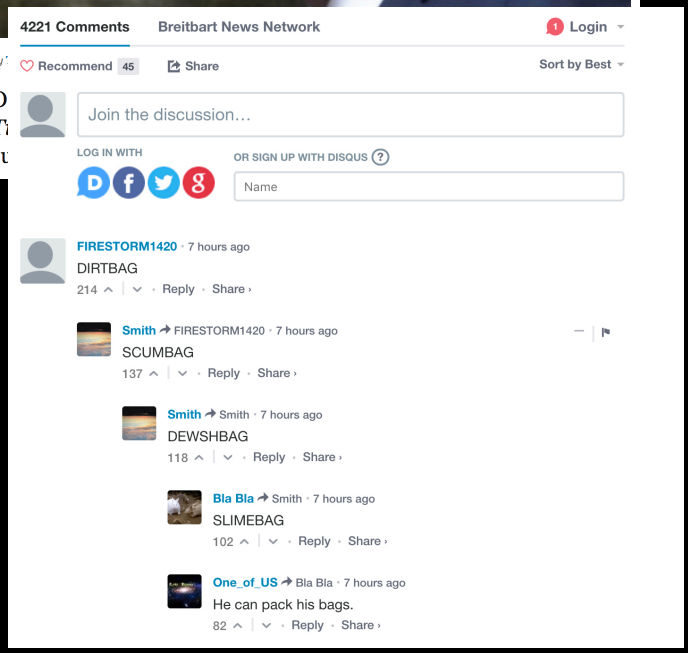
Pro: data collection, sentiment anaylsis, external validity

Con: selection, endogeneity

Approach 2: Experiment

1. Construct your own facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. **Random assignment to treatments**





Approach 1: Harvesting

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

Pro: data collection, sentiment anaylsis, external validity

Con: selection, endogeneity

Approach 2: Experiment

1. Construct your own facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. **Random assignment to treatments**

Pro: Identification

Con: int. validity

Social Media in the Sandbox


TITAN
98203

Preliminary pictures (N=200)

Online survey: selection of pictures" (N=9)

Online comments collection

clickworker

Comments classification

Countercomments

clickworker

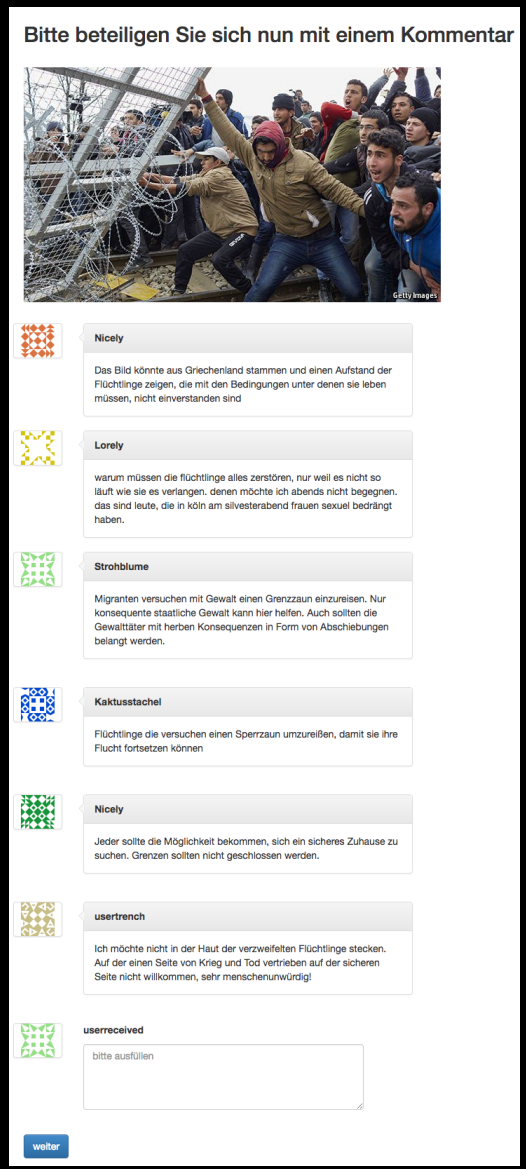
Treatments construction

ONLINE EXPERIMENT

clickworker

Comments classification

Creating an online dicussion forum on social topics

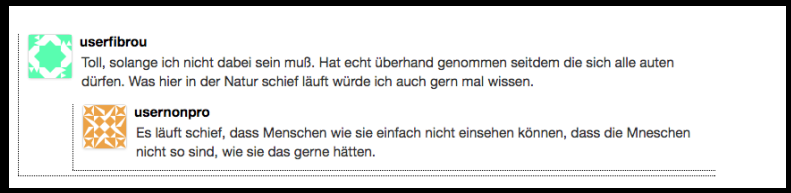


- 9 different pictures from 4 different topics
- Users remain anonymous and are given an avatar and an username to use them in the discussion
- 6 comments are displayed: a mix of neutral, positive/friendly and negative/hostile comments (Example comments)

Censoring treatment

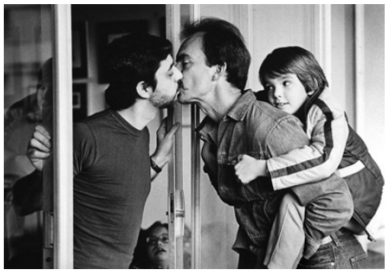
- The negative/hostile comments are deleted
- Two treatments:
  - censored
  - extremely censored

Counterspeaking treatment



- Same pictures and same pool of comments
- Hostile comments are now countercommented

Rating of comments



username

Auch homosexuelle Paare können gute Eltern sein.

Ist der Kommentar freundlich oder feindselig gegenüber der im Foto dargestellten Gruppe? :

sehr freundlich 1 2 3 4 5 6 7 8 9 sehr feindselig

nicht zu bewerten

Welche Merkmale treffen auf den Kommentar zu?

- **Beinhaltet negative Vorurteile**
- **Nutzt rassistische Beleidigungen**
- **Beinhaltet beleidigende, erniedrigende oder abwertende Worte**
- **Ruft zu Gewalt, Drohungen oder Diskriminierung auf**
- **Nutzt sexistische Beleidigungen**
- **Die sexuelle Orientierung oder das Geschlecht/Gender wird herabgesetzt oder stigmatisiert**

Weiter

Rating of comments

Hate speech score:

Is the comment friendly or hostile towards the group represented in the picture? *(Indicate from 1 to 9 where 1 means very friendly and 9 means very hostile)*

Rating of comments

Hate speech score:

Is the comment friendly or hostile towards the group represented in the picture? *(Indicate from 1 to 9 where 1 means very friendly and 9 means very hostile)*

Hate Speech Indicator:

Which of the characteristics applies to the comment?

- *negative stereotypes*
- *racist slurs*
- *demeaning language*
- *encourages violence*
- *sexist slurs*
- *stigmatizes gender or sexual orientation*

Results

Hate speech score by topic and treatment

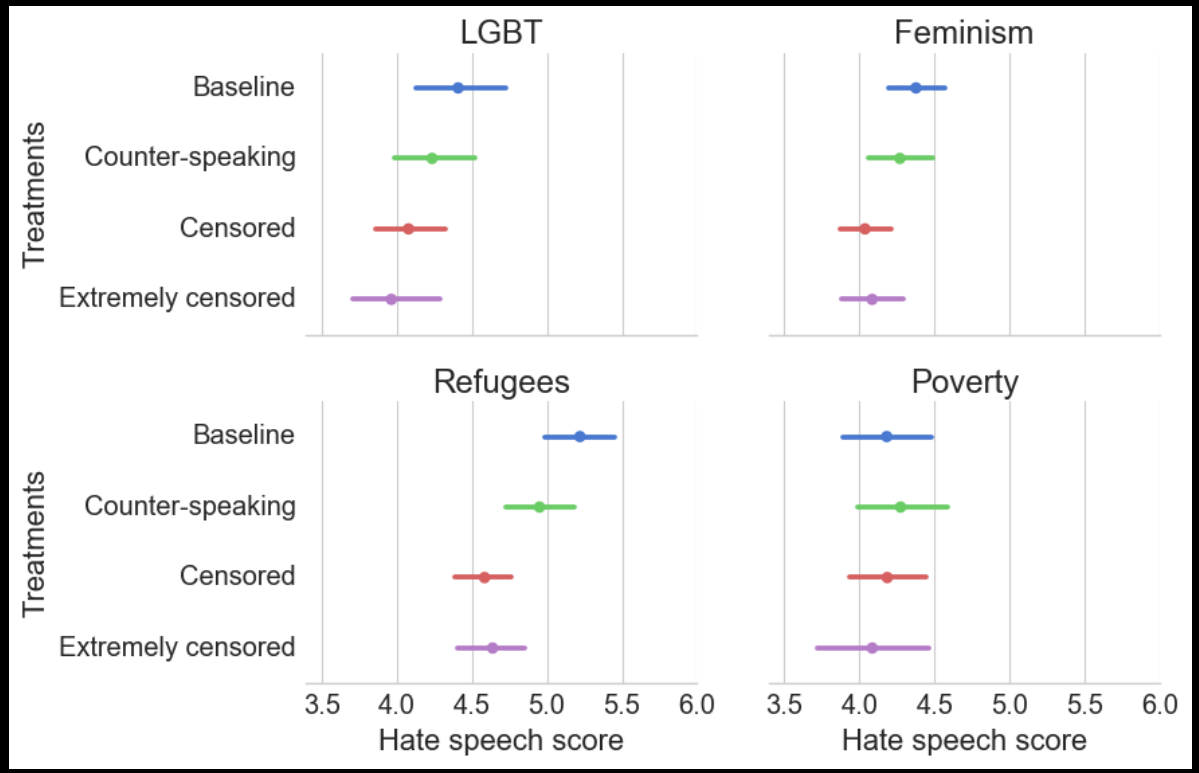


Figure: Average hate speech score across treatments and topics

Distribution of hate speech score

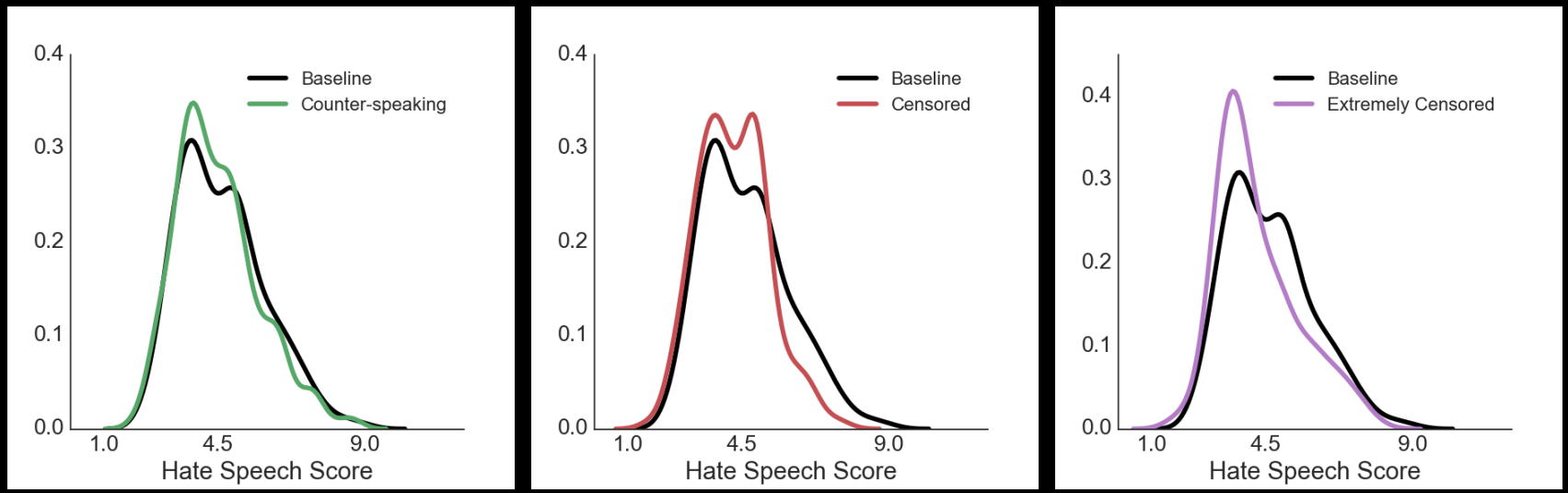


Figure: Density distribution of average hate speech score across treatments. The graph depicts the 1 to 9 score scale

Participants are slightly more prone to use strongly hateful language in the extremely censored treatment.

Hate Speech Indicator

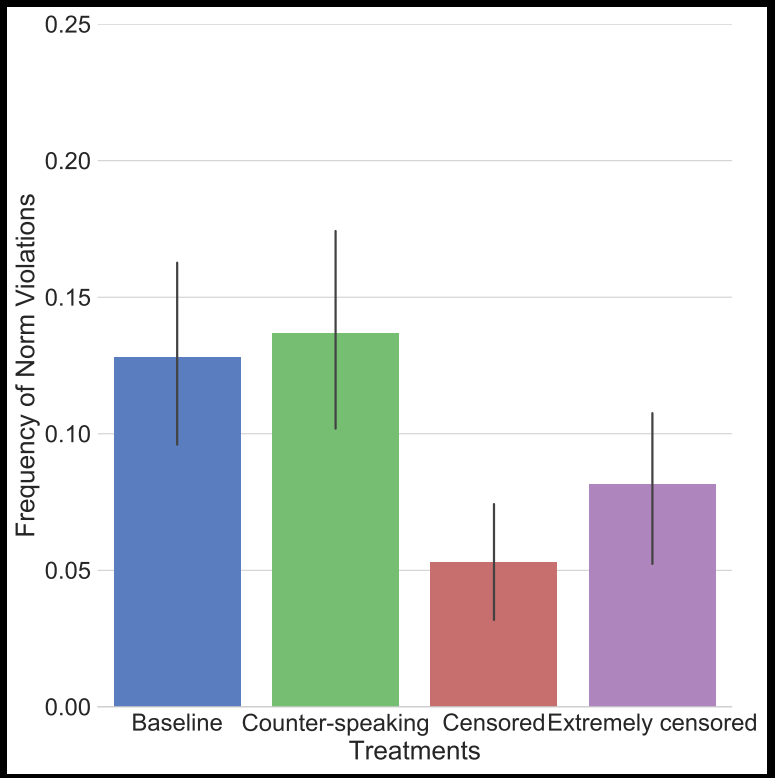


Figure: Proportion of comments that were labeled as hate speech across treatments. Error bars at 95% CI

Proportion of participants

Contributions

This project represents a step forward in the research on online hate speech. The results can help **design better informed interventions** to tackle online hate speech using social norms perception as means for normative change.

Our results add to the literature of social norms, it presents **empirical evidence** of the **effect of social norm perception on the willingness to engage in online hate speech**, even in anonymous contexts, without direct punishment, and controlling for selection effects.

Spillover Effects of Hate Speech: A Natural Experiment


Bild

Motivation

Do external events such as terrorist attacks lead to changed social norms regarding the expression of hateful views?

External shocks effects on hate expression

A spread of hostile and hate expression is normally linked to terrorist attacks:

- Public expression of anti-foreign sentiment following attacks (Legewie, 2013; Hanes and Machin, 2014).
- Increase in hate speech in online social media context (Williams and Burnap, 2015)

Breakdown of modesty norms towards may spill over to modesty norms towards other minority groups. Spillovers: Increases in public expression of hate against some minorities might also spur increases in expression of hate against other minorities (CSBS, 2017).

Change in norms?

External shocks effects on hate expression: mechanisms

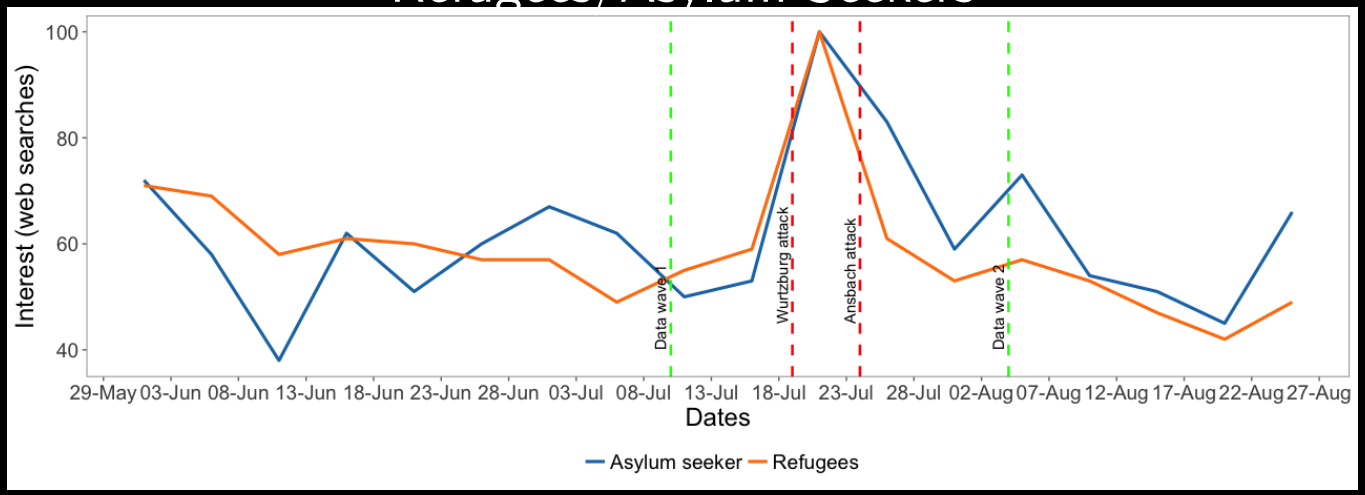
Individuals conform to salient norms (Bicchieri, 2005; Cialdini and Goldstein, 2004) and avoid expressing unpopular opinions (Bursztyn et al., 2016).

Individual's perception of **social acceptability** of hate speech affects willingness to express it publicly (Bursztyn et al., 2017).

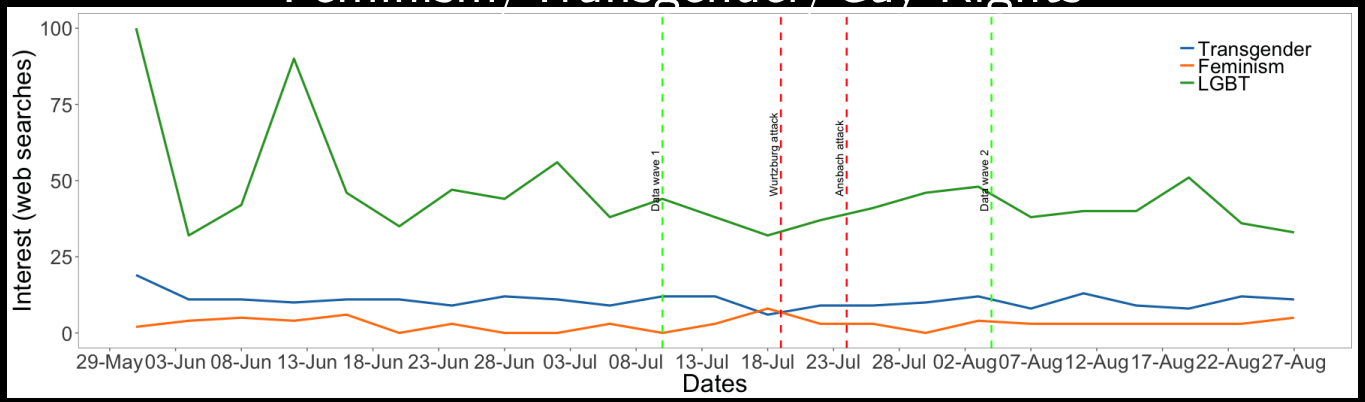
External shocks (e.g. reactions to terrorist attacks...) might induce changes in the social acceptability of certain (extreme) opinions and in the likelihood that these opinions are publicly expressed.

Attack associated with refugees

Refugees/Asylum Seekers



Feminism/Transgender/Gay Rights



Data selection

- We selected data from the baseline and the censored treatments
- We selected data from the refugees, feminism and LGBT comments threads. The feminist and LGBT comments were merged in a new "Other" category.
- The comments were rated by 577 different raters. Raters were asked to rate 30 comment

	Wave 1		Wave 2	
Treatment	Refugees	Other	Refugees	Other
Baseline	135	227	135	228
Censored	136	225	135	226
Extremely censored	134	225	123	204
Total	405	677	393	658

Results

Change in mean hate speech score: baseline treatment

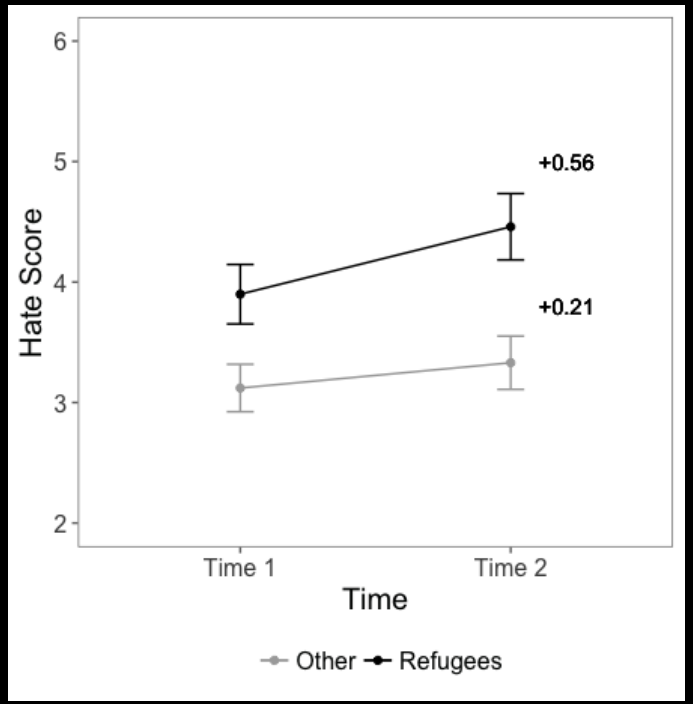


Figure: Error bars at 95%. The graph displays the changes in hate speech score before and after the terrorist attacks for the two categories.Error bars at 95%.

Change in mean hate speech score

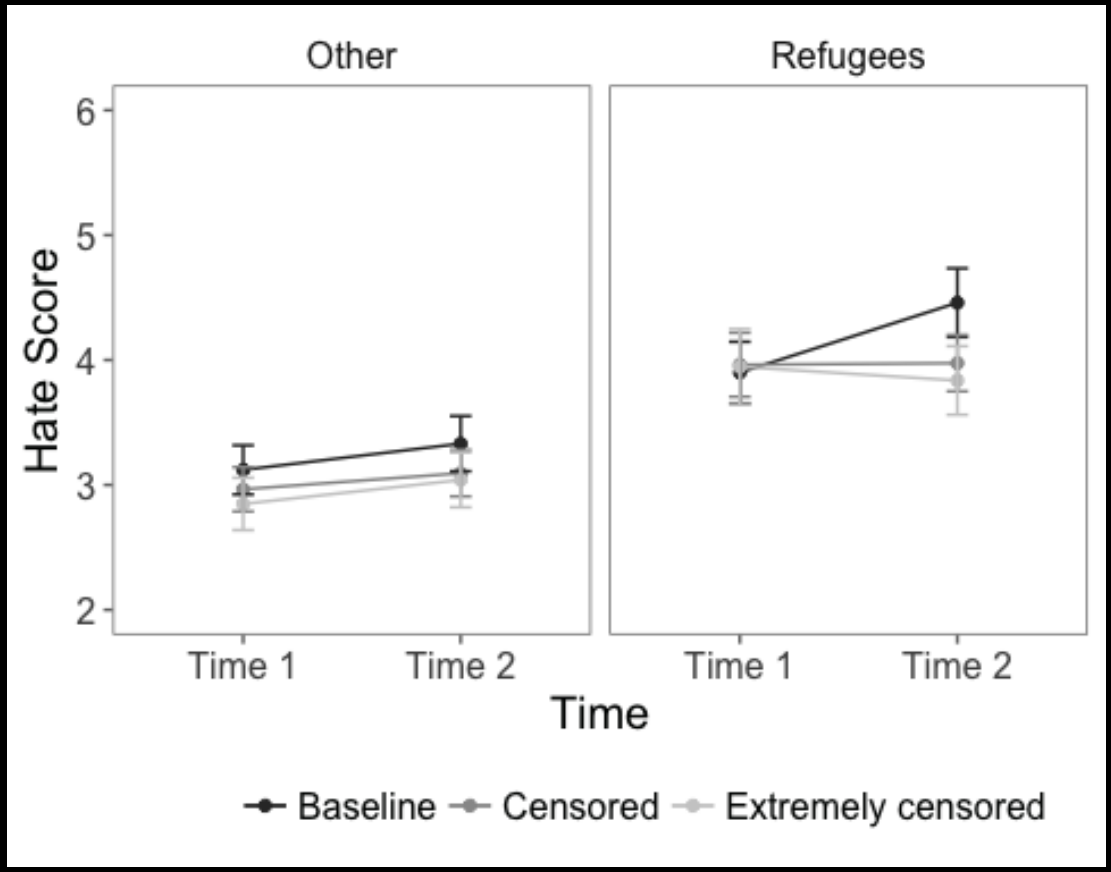
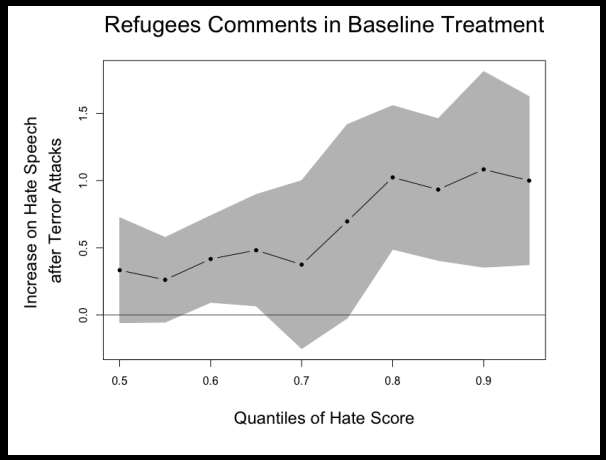
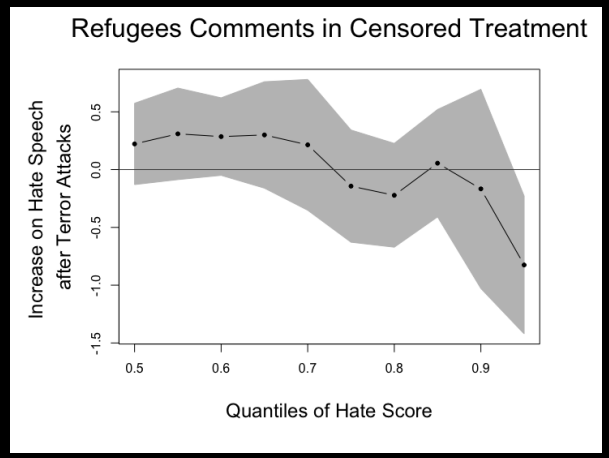
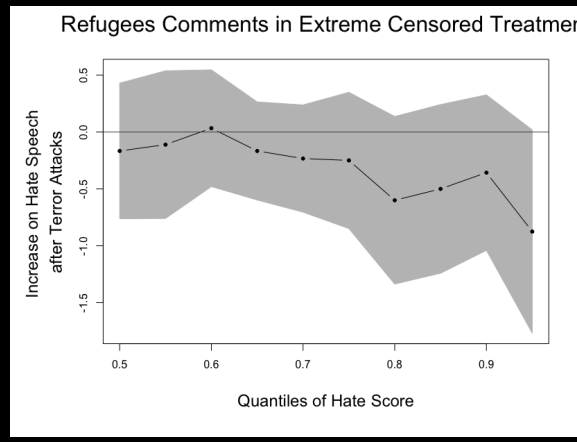


Figure: Error bars at 95%. Left: changes in hate speech score before and after the terrorist attacks for the other category. Right: refugees.

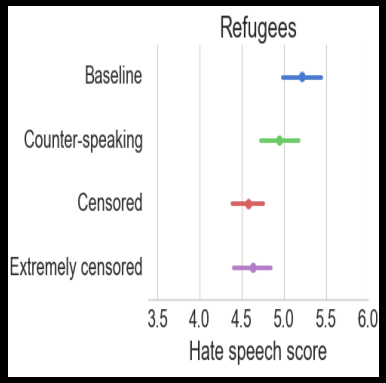
Increase of hate speech towards refugees



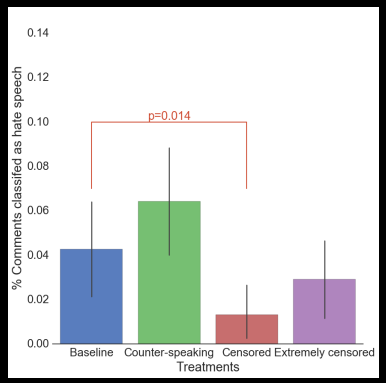




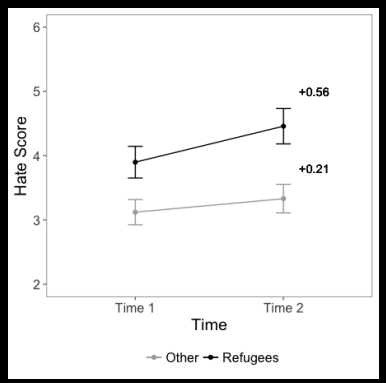
Normative change and culture of hate: Summary



- Hate speech is context dependent:
Descriptive norms matter.
- Censoring extremely negative comments leads to
less hate speech.



- Extreme Censoring may lead to
more extreme comments.



- External shocks may erode hate speech norms.
- Weak spillover of hate speech to **unrelated domains**
- Interventions in one domain may
inhibit negative dynamics in other domains

Appendix

Treatment Effects - Study 1

$$Y_{ij} = \beta_0 + \beta_1 Treatment + \beta_2 Topic + u_j + \epsilon_{ij}$$

	Hate speech score:	
	Model (1)	Model (2)
Constant	4.61 (0.11)**	4.41 (0.13)**
Counter-speaking	−0.13 (0.16)	−0.14 (0.15)
Censored	−0.38 (0.16)*	−0.39 (0.15)*
Extremely censored	−0.38 (0.16)*	−0.40 (0.16)*
LGBT		−0.00 (0.09)
Refugees		0.65 (0.09)**
Feminism		0.03 (0.09)
Observations	1469	1469
Number of Subjects	180	180

Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$

Table: Results from multilevel random models of hate speech score. Model 1 shows main effects of treatments. Model 2 shows main effects of treatments and topics. The baseline treatment and the topic poverty serve as the reference categories.

Non-parametric results

Increase in hate speech score in time 2 by topic

	Other	Refugees
Constant	3.11 (0.16)**	3.90 (0.17)**
Time 2	0.24 (0.23)	0.56 (0.25)*
Obs.	455	270
Groups:Subjects	94	90
Var: Subjects	0.91	0.86

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score

Increase in hate speech score by topic

	Model 1
Constant	3.11 (0.15)**
Time 2	0.24 (0.22)
Refugees	0.78 (0.14)**
Time*Refugees	0.36 (0.20)·
Obs.	725
Groups:Subjects	94
Var:Subjects	0.75

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score

Interaction with experimental treatments

	Model 1
Main effects	
Constant	3.11 (0.15)**
Refugees	0.78 (0.14)**
Time 2	0.24 (0.21)
Censored	−0.14 (0.21)
Extremely censored	−0.24 (0.21)
Interaction effects	
Refugees*Time 2	0.36 (0.20)·
Refugees*Censored	0.22 (0.20)
Refugees*Extremely Censored	0.32 (0.20)
Time 2*Censored	−0.11 (0.30)
Time 2*Extremely Censored	−0.05 (0.31)
Refugees*Time 2* Censored	−0.48 (0.28)·
Refugees*Time 2* Ext. Censored	−0.66 (0.28)*
Obs.	2133
Groups:Subjects	274
Var: Subjects	0.71

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score by treatment, time and category

Comments example

Friendly:

- ,,Ich wünsche mir, für jeden der hier leben und arbeiten will, dass er alle erdenkliche Hilfe bekommt um es für sich und seine Familie realisieren zu können" (*Refugees*)
- ,,Super Daumen hoch für diese Leute die den Mumm haben sich der Ignoranz zu stellen" (*Feminism*)

Neutral:

- ,,Generell bin ich dagegen sich in der Öffentlichkeit wild zu küssen. Aber gegen einen Kuss habe ich nichts" (*LGBT rights*)
- ,,Letzten Endes lediglich ein gewöhnungsbedürftiger Anblick. Bis auf die Gesichtsverschleierung kommt es einer Nonne fast schon gleich" (*Refugees*)

Hostile:

- ,,Einfach nur absurd, dass unsere Politiker all diese Leute einfach ohne Papieren einreisen lassen. Die meisten sind eine Gefahr für uns und unsere Kinder. Und können dann nichtmal abgeschoben oder bestraft werden ... "(*Refugees*)
- ,,Meine Toleranz hat Grenzen. Transsexuelle haben beim Militär nichts zu suchen" (*Feminism*)

Kernel density estimates

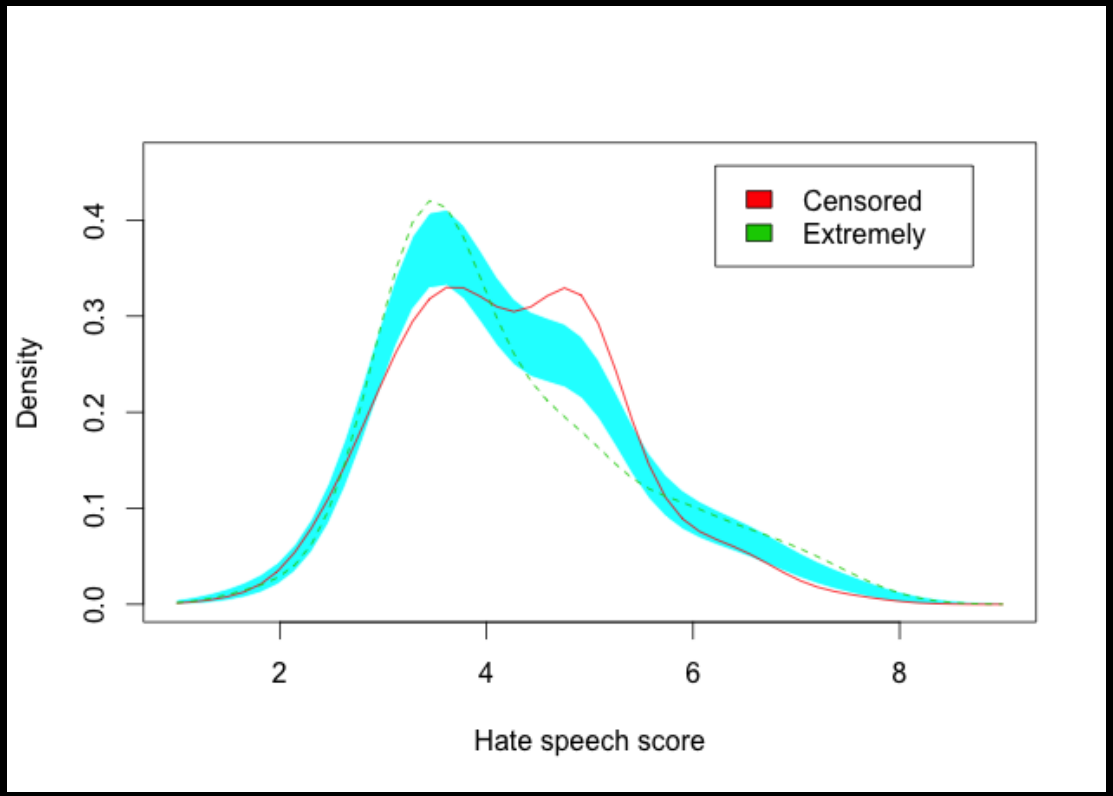


Figure: Kernel density estimates for the hate speech score in censored(green line) and extremely censored (blue line) treatments

Non-parametric results

Treatment	Mean(Score)	Median(Score)	p vs baseline
Baseline	4.61 (1.30)	4.33	
Counter-speaking	4.48 (1.26)	4.33	0.5483
Censored	4.24 (1.06)	4	0.0023
Extremely censored	4.24 (1.24)	4	0.0000
Total	4.39	4.33	

Table: Mean and median hate speech score in the different treatments (standard deviation in parentheses). In column 4, the level of hate speech in the treatments is compared to the baseline level (Kruskal-Wallis test).

parametric results

Models

$$Y_{ij} = \beta_0 + \beta_1 \mathit{Treatment} + u_j + \epsilon_{ij} \tag{1}$$

$$Y_{ij} = \beta_0 + \beta_1 \mathit{Treatment} + \beta_2 \mathit{Topic} + u_j + \epsilon_{ij} \tag{2}$$

$$Y_{ij} = \beta_0 + \beta_1 \mathit{Topic} + \beta_2 (\mathit{Treatment} * \mathit{Topic}) + u_j + \epsilon_{ij} \tag{3}$$

.

Results from multilevel random models of hate speech score

	(1)	(2)	(3)
Main effects			
Constant	4.61 (0.11)**	4.41 (0.13)**	4.20 (0.18)**
Counter-speaking	−0.13 (0.16)	−0.14 (0.15)	
Censored	−0.38 (0.16)*	−0.39 (0.15)*	
Extremely censored	−0.38 (0.16)*	−0.40 (0.16)*	
LGBT		−0.00 (0.09)	0.21 (0.18)
Refugees/Multiculturality		0.65 (0.09)**	1.01 (0.18)**
Feminism		0.03 (0.09)	0.19 (0.17)
Interaction effects			
Poverty*Counter-speaking			0.08 (0.25)
Poverty*Censored			−0.02 (0.25)
Poverty*Extremely			−0.12 (0.26)
LGBT*Counter-speaking			−0.16 (0.20)
LGBT*Censored			−0.33 (0.20)
LGBT*Extremely			−0.44 (0.21)*
Refugees*Counter-speaking			−0.26 (0.19)
Refugees*Censored			−0.65 (0.19)**
Refugees*Extremely			−0.61 (0.19)**
Feminism*Counter-speaking			−0.11 (0.18)
Feminism*Censored			−0.34 (0.18)†
Feminism*Extremely			−0.28 (0.19)
Random Parts			
Groups: Subjects	180	180	180
Var: Subjects	0.43	0.43	0.43
Residual Variance	1.06	0.97	0.97
Obs.	1469	1469	1469

Hate Speech Indicator by participant

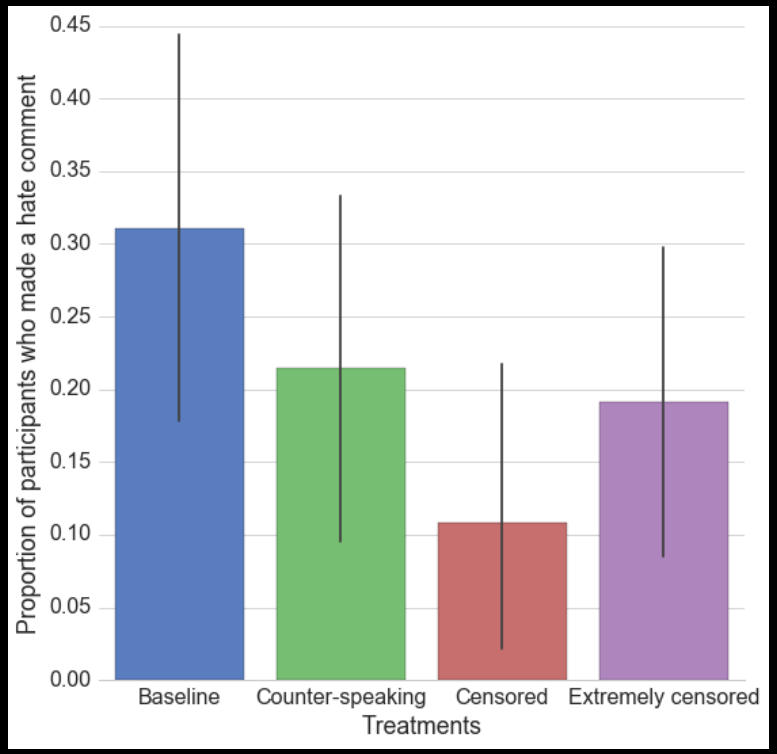


Figure: Error bars at 95%. Left: Number of participants that made at lest one hate comment.

back

References I

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, Cambridge.

Blanchard, F. A., Crandall, C. S., Brigham, J. C., and Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, 79(6):993.

Bursztyn, L., Callen, M., Ferman, B., Gulzar, S., Hasanain, A., and Yuchtman, N. (2016). Political identity: Experimental evidence on anti-americanism in pakistan.

Bursztyn, L., Egorov, G., and Fiorin, S. (2017). From extreme to mainstream: How social norms unravel. Working Paper 23415, National Bureau of Economic Research.

Center for the Study of Hate and Extremism (2017). Special status report: Hate crime special status report of hate crime in metropolitan areas. Technical report, California State University, San Bernardino.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.

References II

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In Gilbert, D. T., Fiske, S. T., and Lindzey, G., editors, *The handbook of social psychology*, volume 2, pages 151–192. McGraw-Hill, New York, 4 edition.

Hanes, E. and Machin, S. (2014). Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, 30(3):247–267.

Legewie, J. (2013). Terrorist events and attitudes toward immigrants: A natural experiment. *American Journal of Sociology*, 118(5):1199–1245.

Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, pages 1–21.

Stangor, C., Sechrist, G. B., and Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, 27(4):486–496.

Williams, M. L. and Burnap, P. (2015). Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238.