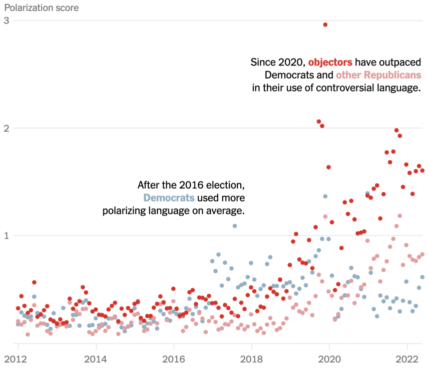
Do Sanctioning and Monitoring Affect Political Elites' Online Toxicity?

Evidence from a Field Experiment on US General Election Candidates

Yunus Emre Orhan (NDSU)
Val Mechkova (U. Gothenburg)
Dan Pemstein (NDSU)
Brigitte Seim (UNC)
Steven Wilson (Brandeis)

Motivation

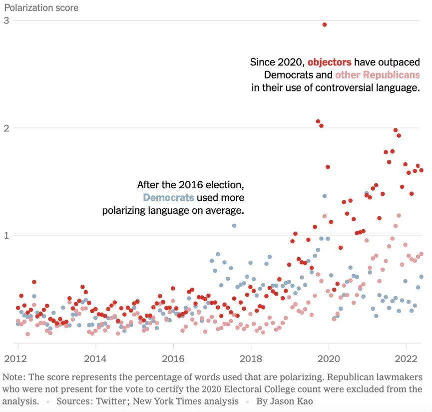
- Increasing online polarization and toxicity

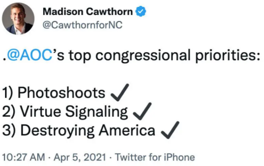


Note: The score represents the percentage of words used that are polarizing. Republican lawmakers who were not present for the vote to certify the 2020 Electoral College count were excluded from the analysis. • Sources: Twitter; New York Times analysis • By Jason Kao

Motivation

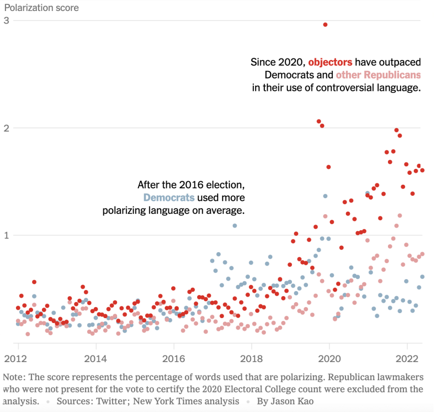
- Increasing online polarization and toxicity

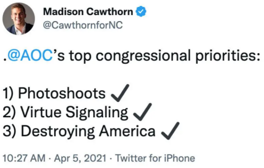




Motivation

- Increasing online polarization and toxicity





Motivation

- Increasing online polarization and toxicity
- Potential pernicious effects of elites' toxic language
  - Swamping out constructive debate (Druckman, Peterson & Slothuus 2013)
  - Exacerbating polarization (Bail et al. 2018)
  - Encouraging harassment of historically under-represented groups (Mechkova & Wilson 2021)
  - Stimulating political violence (Feuer, Schmidt & Broadwater 2022)

Motivation

- Increasing online polarization and toxicity
- Potential pernicious effects of elites' toxic language
  - Swamping out constructive debate (Druckman, Peterson & Slothuus 2013)
  - Exacerbating polarization (Bail et al. 2018)
  - Encouraging harassment of historically under-represented groups (Mechkova & Wilson 2021)
  - Stimulating political violence (Feuer, Schmidt & Broadwater 2022)
- Can we *cheaply* reduce elites' online toxicity?

Research Questions

1. Does bottom-up social sanctioning reduce politicians' online toxic speech?
   (Rasinski and Czopp, 2010; Munger 2017)
2. Does top-down monitoring reduce politicians' online toxic speech?
   (Grossman and Hanlon, 2014; Grossman and Michelitch, 2018; Nyhan and Reifler, 2015)
3. Does top-down monitoring reduce politicians' willingness to communicate online (chill speech)?

Contributions

- Adaptation of bottom-up sanctioning to elite targets
- Application of monitoring intervention to online behavior
- Comparative analysis of political elite social media behavior

  including Tunisia, Turkey, Brazil, India, Italy, the Philippines, France, and Australia.

Data

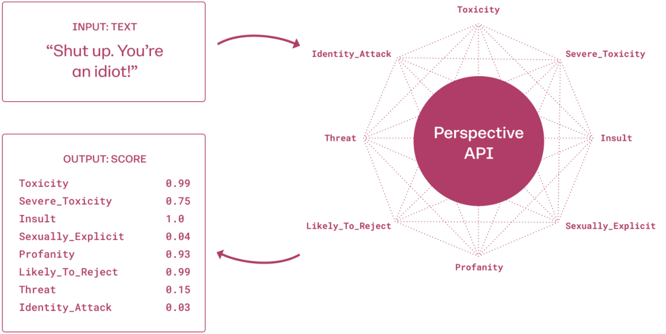
- Country: The US
- Period: Oct-13-2022 - Dec-15-2022
- Data: 2022 Midterm Election Candidates
- Sample size: 3560
- Inclusion Criteria:
  1. General election candidate for legislative office (Fed/State)
  2. Affiliation with Democratic or Republican Party
  3. Twitter account

Hypotheses

- H1: Top-down monitoring will decrease toxicity
- H2: Bottom-up social sanctioning will decrease toxicity
- H3: Top-down monitoring will decrease toxicity more than bottom-up social sanctioning
- H4: Top-down monitoring will decrease tweet frequency (number of tweets per week)

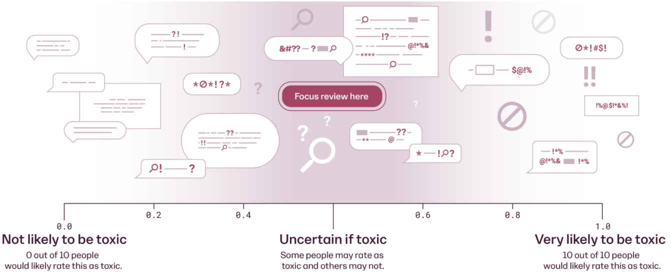
Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)



Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)



Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

Toxicity Level	Description of level
Very Toxic	A comment that is very hateful, aggressive, disrespectful, or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
Toxic	A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.
Not Toxic	A neutral, civil, or even nice comment very unlikely to discourage the conversation.
I'm not sure	The comment could be interpreted as toxic depending on the context but you are not sure.

Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

Category	**Definition**
Profanity/ Obscenity	Swear words, curse words, or other obscene or profane language.
Identity-based negativity	A negative, discriminatory, stereotype, or hateful comment against a group of people based on criteria including (but not limited to) race or ethnicity, religion, gender, nationality or citizenship, disability, age, or sexual orientation.
Insults	Inflammatory, insulting, or negative language towards a person or a group of people. Such comments are **not** necessarily identity specific.
Threatening	Language that is threatening or encouraging violence or harm, including self-harm.

Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

COMMENT

You're a real idiot, you know that.

☐ This comment is not in English or is not human-readable.

Rate the toxicity of this comment.

Very toxic: A comment that is very hateful, aggressive, disrespectful, or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.

Toxic: A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.

○ Very toxic
○ Toxic
○ Maybe, not sure
○ Not Toxic

Does this comment contain obscene or profane language?

Profanity/obscenity: Swear words, curse words, or other obscene or profane language.

○ Yes
○ Maybe, not sure
○ No

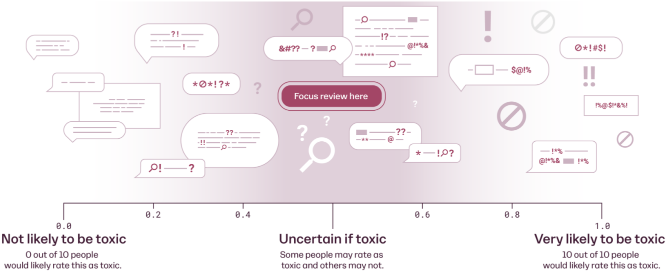
Does this comment contain identity-based negativity?

Identity-based negativity: A negative, discriminatory, stereotype, or hateful comment against a group of people based on criteria including (but not limited to) race or ethnicity, religion, gender, nationality or citizenship, disability, age, or sexual orientation.

○ Yes
○ Maybe, not sure
○ No

Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)



Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)
  - TweetCount: log(# of tweets in the Pre/Post-periods + 1).

Outcomes and Controls

- Dependent Variables / Outcomes
  - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)
  - TweetCount: log(# of tweets in the Pre/Post-periods + 1).
- Control Variables
  - Candidate Level
    - Overall Toxicity (Candidate toxicity bin)
    - Party ID, Sex, Incumbency, Twitter Account Type
  - Election Level
    - State, District, Type, District Competitiveness
  - Treatment Level
    - Sanctioning Timing and Group
    - Monitoring Date
    - Other Arm

Experimental Conditions

1. Top-Down Monitoring Condition
   (Grossman and Hanlon, 2014; Grossman and Michelitch, 2018; Nyhan and Reifler, 2015)
2. Bottom-up Social Sanctioning Condition
   (Munger 2017)
3. Control Condition

Top-Down Monitoring Arm

*Dear [$***Candidate Full Name***].*

We are two independent researchers at North Dakota State University. We are not affiliated with any partisan group in any way.

*We are writing to let you know we are conducting research on the use of toxic language on Twitter by candidates, specifically how use of such language affects election outcomes. We are monitoring your Twitter account [***@handle(s)***] and will compile your tweets that use toxic language. Just before the election, we will write a post on the Monkey Cage blog of the Washington Post that discusses our findings regarding patterns in the use of toxic language.*

Sincerely, Drs. Daniel Pemstein and Yunus Orhan

Bottom-Up Social Sanctioning Arm

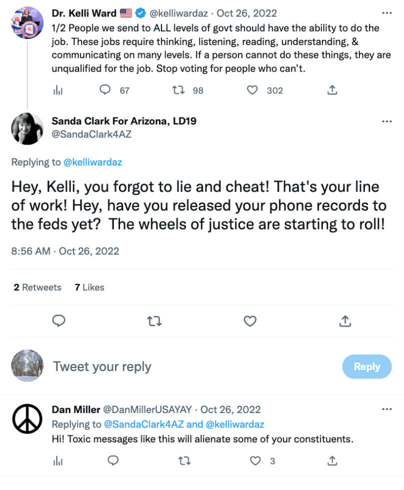
Mention	**Preamble** (1 of 5)	**Text** (1 of 5)
	Good day!	Just remember that some of your constituents may be upset by toxic messages like this.
	Hi!	This toxicity is upsetting to some of your constituents.
@[Harasser Candidate]	Hey there.	Some of your constituents are going to be upset by this toxic message.
	Hi there.	That is a toxic thing to say, and it will push away some of your constituents.
	Hello!	Toxic messages like this will alienate some of your constituents.

Progressive Pictures & Banners

Conservative Pictures & Banners

Brian Smith
News junkie
Beer drinker

Sanctioning Example



Sanctioning Example

Sanctioning Criteria

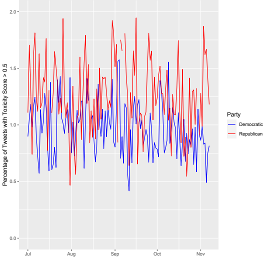
- The Google's Jigsaw Perspective (GJP) scores are imperfect proxies for sanctionable content.
- Hybrid Procedure: Combining machine coding and human discretion

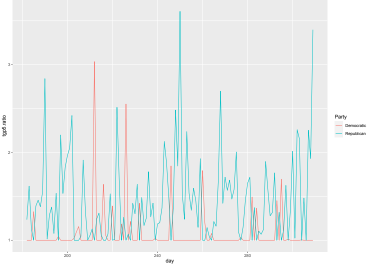
Estimands

- Intention to treat (ITT) effect
  - Differences-in-differences (DiD) design
  - Pre/post treatment periods (1 week)
  - Quantity of interest is interaction between condition and period
- Total effect of treatment on the treated (TOT)
  - Sanctioning: only 121 actual sanction events
  - Monitoring: a handful of bounced emails
  - DiD, with instrumental variable (experimental condition)

Descriptive Findings

1. Candidates rarely engage in stark toxicity on Twitter
2. Democrats generate more toxic tweets
3. Republicans' tweets are more likely to be toxic than Democrats' tweets
4. Men are twice more likely than women to post toxic tweets





Top-Down Monitoring Intervention Results

	ITT				TOT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.097***	-29.8	0.135***	-51.4**	0.097***	-30.3	0.135***	-52.0**
	(0.003)	(18.8)	(0.004)	(24.2)	(0.003)	(18.9)	(0.004)	(24.2)
Treatment	0.011**	0.011***	0.013**	0.014***	0.011**	0.012***	0.014**	0.015***
	(0.005)	(0.004)	(0.006)	(0.005)	(0.005)	(0.004)	(0.006)	(0.005)
Period	0.015***	0.015***	0.017***	0.017***	0.015***	0.015***	0.017***	0.017***
	(0.005)	(0.004)	(0.006)	(0.005)	(0.005)	(0.004)	(0.006)	(0.005)
Treatment × **Period**	-0.014**	-0.014***	-0.015*	-0.017**	-0.015**	-0.015**	-0.016*	-0.018**
	(0.007)	(0.005)	(0.008)	(0.007)	(0.007)	(0.006)	(0.009)	(0.007)
Controls	X	✓	X	✓	X	✓	X	✓
0-Tweet	0	0	NA	NA	0	0	NA	NA
R^2	0.002	0.425	0.002	0.391	0.002	0.425	0.002	0.391
Observations	7,120	7,120	5,223	5,223	7,120	7,120	5,223	5,223

- Monitoring reduces P(toxic) by 1.5 points, on average
- This represents a 15% reduction for a typical candidate

Bottom-Up Social Sanctioning Intervention Results

	ITT				TOT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.105***	0.015	0.141***	0.043***	0.105***	0.013	0.141***	0.048***
	(0.003)	(0.011)	(0.004)	(0.015)	(0.003)	(0.011)	(0.004)	(0.015)
Treatment	-0.003	-0.006	0.004	0.003	-0.050	-0.058	0.061	0.031
	(0.005)	(0.004)	(0.006)	(0.005)	(0.072)	(0.046)	(0.089)	(0.059)
Period	-0.003	-0.003	-0.002	-0.002	-0.003	-0.001	-0.002	-0.003
	(0.005)	(0.004)	(0.006)	(0.005)	(0.005)	(0.003)	(0.006)	(0.004)
Treatment × **Period**	0.011	0.011**	0.003	0.004	0.163	0.104**	0.047	0.084
	(0.007)	(0.005)	(0.009)	(0.007)	(0.102)	(0.049)	(0.126)	(0.061)
Controls	X	✓	X	✓	X	✓	X	✓
0-Tweet	0	0	NA	NA	0	0	NA	NA
R^2	0.0005	0.401	0.0004	0.371	0.0005	0.401	0.0004	0.371
Observations	7,120	7,120	5,184	5,184	7,120	7,120	5,184	5,184

- Sanctioning increases P(toxic) by 10 points, on average
- This represents a 100% increase for a typical candidate

Chilling Effect?

	ITT		TOT	
	(1)	(2)	(3)	(4)
Constant	1.63***	-168.4	1.63***	-164.2
	(0.029)	(158.2)	(0.029)	(158.3)
Treatment	-0.013	-0.012	-0.014	-0.013
	(0.042)	(0.031)	(0.045)	(0.034)
Period	0.077*	0.077**	0.077*	0.077**
	(0.042)	(0.031)	(0.042)	(0.031)
Treatment × **Period**	-0.051	-0.051	-0.055	-0.055
	(0.059)	(0.044)	(0.064)	(0.047)
Controls	X	✓	X	✓
R^2	0.0008	0.447	0.0008	0.447
Observations	7,120	7,120	7,120	7,120

- No evidence that candidates tweet less when monitored

Key Takeaways

1. Top-down monitoring REDUCES toxicity among elites
2. Bottom-up sanctioning INCREASES toxicity among elites
3. Top-down monitoring HAS NO EFFECT on candidate tweet volume

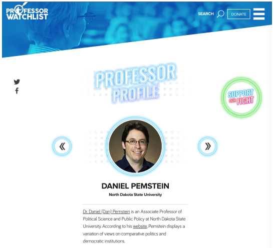
Key Takeaways

1. Top-down monitoring REDUCES toxicity among elites
2. Bottom-up sanctioning INCREASES toxicity among elites
3. Top-down monitoring HAS NO EFFECT on candidate tweet volume

What's next?
Building our own metaketa
Better monitoring interventions
Custom toxicity measures

Key Takeaways

1. Top-down monitoring REDUCES toxicity among elites
2. Bottom-up sanctioning INCREASES toxicity among elites
3. Top-down monitoring HAS NO EFFECT on candidate tweet volume



What's next?
Building our own metaketa
Better monitoring interventions
Custom toxicity measures

Thanks!

RAs:

William Carney, Madison Delorme, Antony Ibrahim, Evan Jones, Maguire Martin, Ruben Orozco, Jonathan Ross, Rohan Tapiawala

Collaborators:




Funding:

National Science Foundation