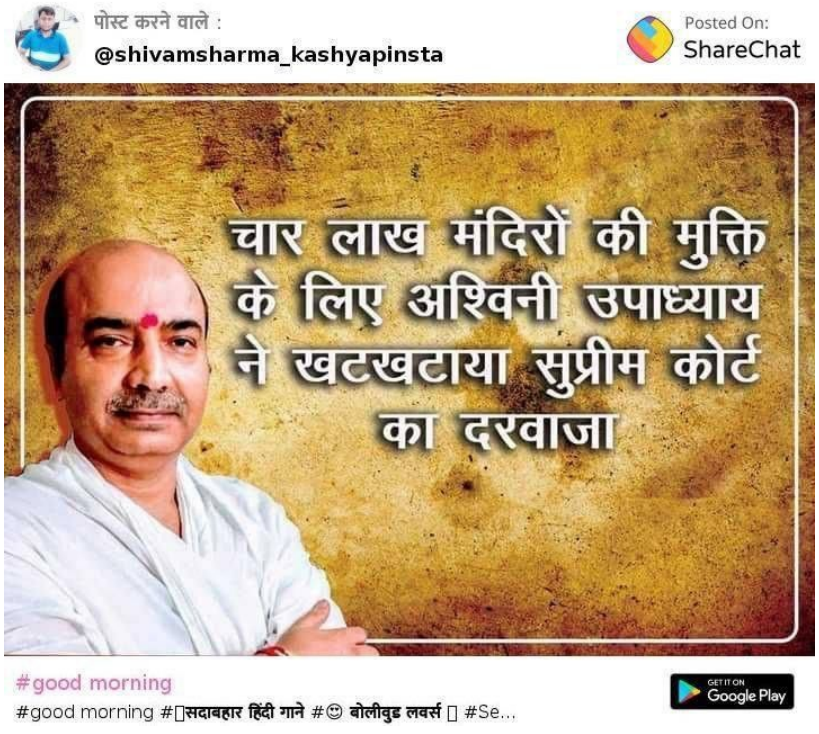
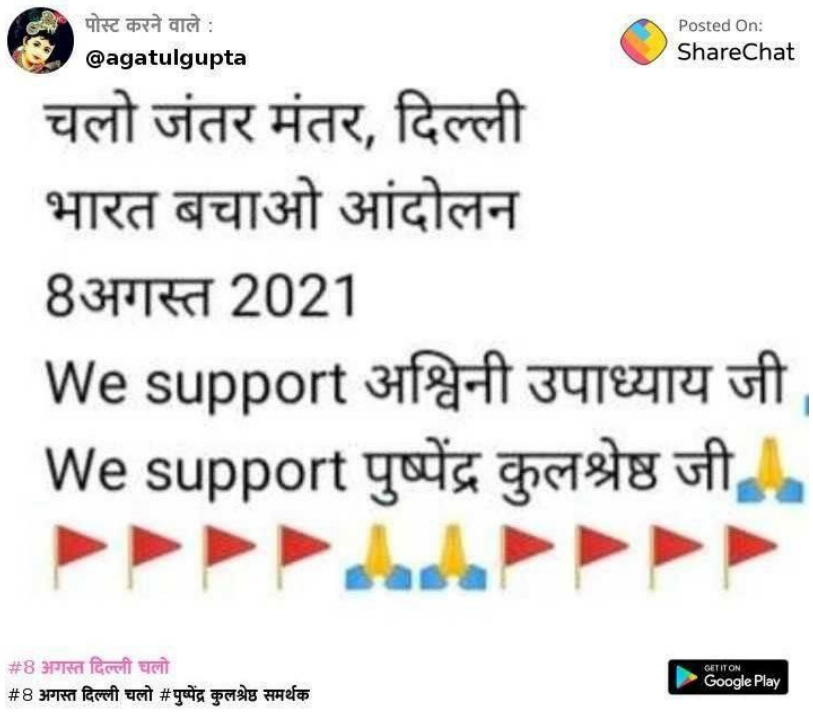
Elections, Leader Identity and Hate Speech

Aarushi Kalra
Brown University

Hate Speech in India





‘Cyber Space is Not Real Space!’

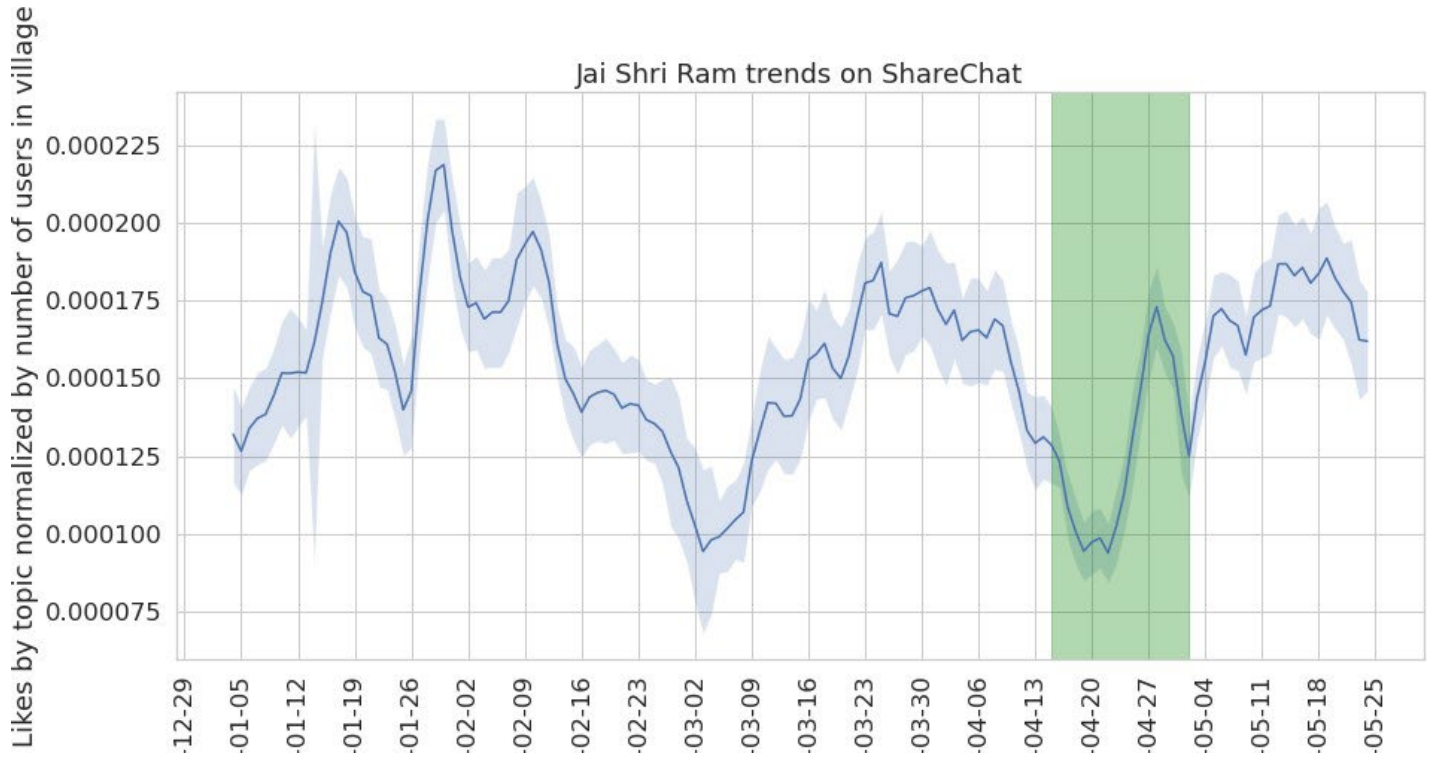
- Mobilizing offline hate through social media

- Posts that target and label certain groups (including Muslims, and human rights activists) incite violence
- Hate Crimes against vulnerable populations incited by Alt-Right groups on Facebook (Müller and Schwarz, 2019)

Hate Speech: Engagement on ShareChat[1]

What *causes* engagement with hateful content on social media?

- Economic Shocks
- Algorithms
- Political Shocks



[1]ShareChat is a content generation with 180 million users in India generating and engaging with content in 14 regional Indian languages.

Motivation: Leader Identity and Hate Speech

- For a given distribution of prejudices: when are anti-minority opinions publicly expressed? (Bursztyn et al., 2020b)
- What are the social 'norms' (Benabou and Tirole, 2011) that incentivize hateful behaviour?
- When do these norms change? (Bursztyn et al., 2020a)
- Role of political leaders in such changes (Meyersson, 2014)

Research Questions

How does expression of anti-minority opinions change with the religious identity of local leaders?

Is hate speech driven by competitive elections where religious identity is salient?

Contributions

Related Literature

Political Polarization:

Gentzkow et al. (2016); Kuziemko and Washington (2018); Boxell et al. (2020)

Norms and Behaviour in Social Networks:

Benabou and Tirole (2011); Halberstam and Knight (2016); Bursztyn et al. (2020b,a)

Leader Identity and People's Behaviour:

Bettinger and Long (2005); Ajzenman et al. (2020); Bhalotra et al. (2021)

Backlash and Populism:

Acemoglu et al. (2013); Mitra and Ray (2014)

Effects of Media, Internet and Social Media:

Gentzkow and Shapiro (2010); Enikolopov et al. (2011); Alcott et al. (2022)

Contributions

- Data in the Wild
  - 180 million users on ShareChat
  - Linked to WhatsApp
  - Differentiate ‘private’ and ‘public’ behavior
  - Hate speech detection in Hindi
- Effect on social media behaviour
- Role of elections
- Role of elected political leaders
- Multi-lingual Hate Speech Classification

Outline

Introduction

Data

Empirical Strategy

Results

Discussion

Data

Background: ShareChat

Bridging the data gap with Indian Content Generation App: ShareChat

- Data in the Wild (2015-now)
- 180 million active monthly users, spending 34 minutes each day on average
- Create and Share Image Content: TikTok ban in July 2020
- Content in 14 non-English regional languages: Focusing on Hindi speaking users in UP
- Particular user base: urban and rural poor in India
- Directly linked to users WhatsApp
- Other forms of engagement on the App
- LatLong locations of users made available to researchers

Politics on ShareChat

Jai Shri Ram

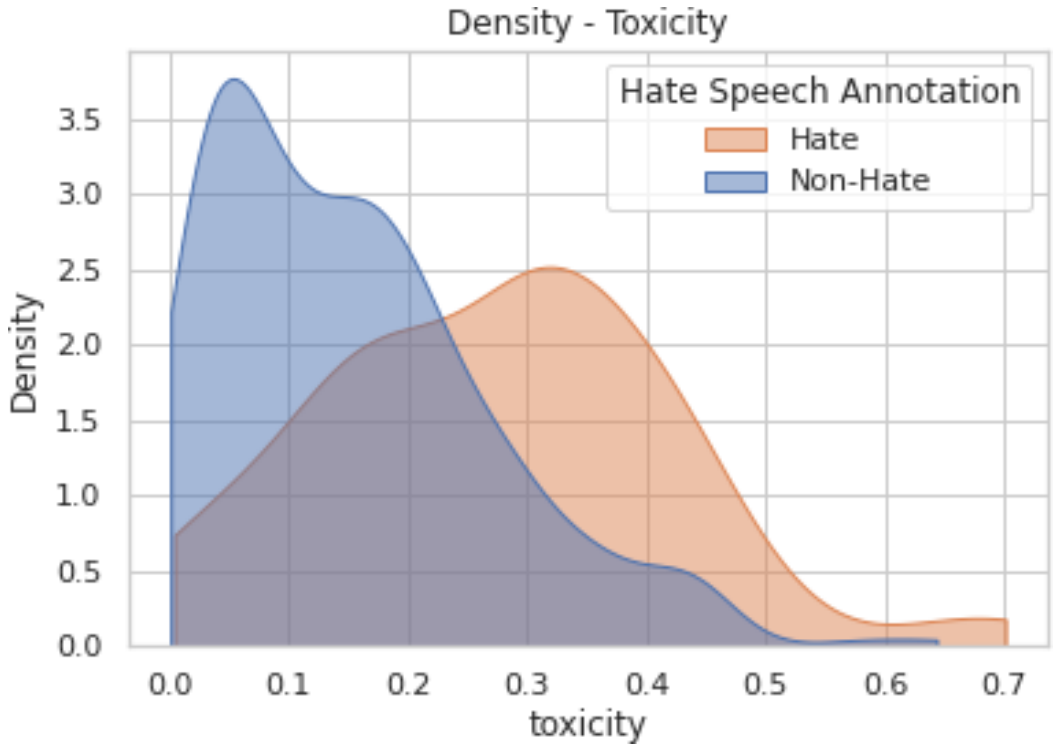
Saffron Nation

Modi

Temple Hindu

Religion Bengal

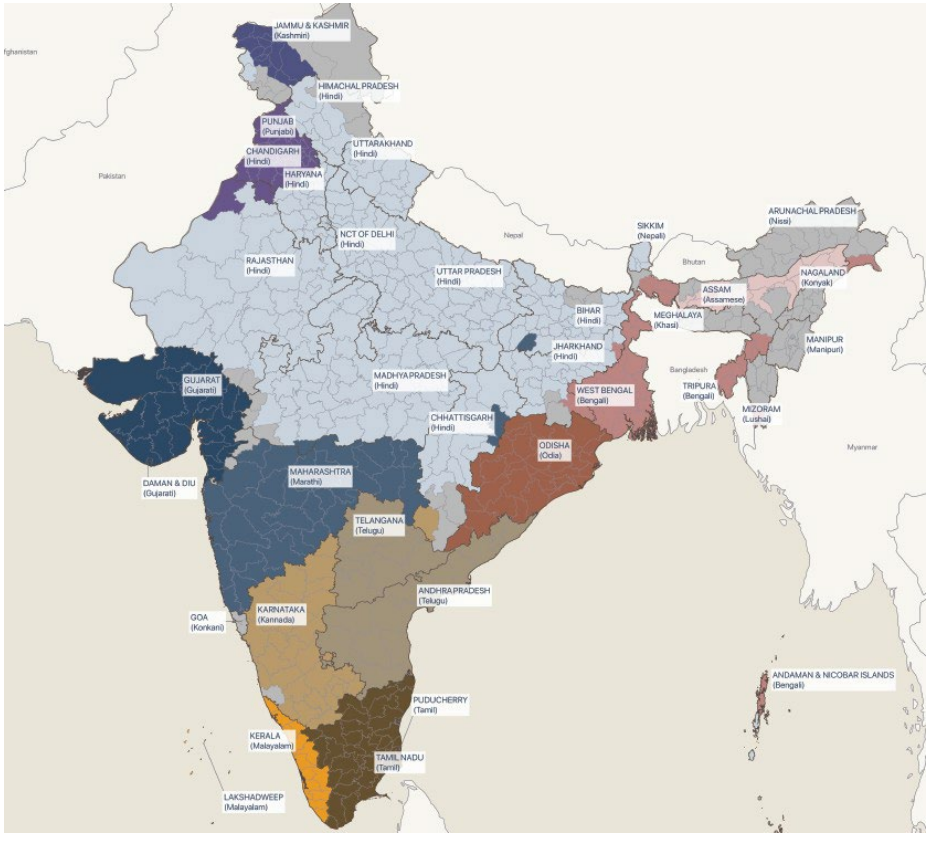
Hateful and Political Posts



Perspective API: Toxicity Scores for Hindi Text Data, by hate annotation

UP Panchayat Elections

- Why Uttar Pradesh (UP)?
- Scraped State Election Commission Website
  - 60,000 village elections in 2015 and 2021
  - Elections in four phases
  - Vote shares of winner and runner up in 2021



Linguistic map of India

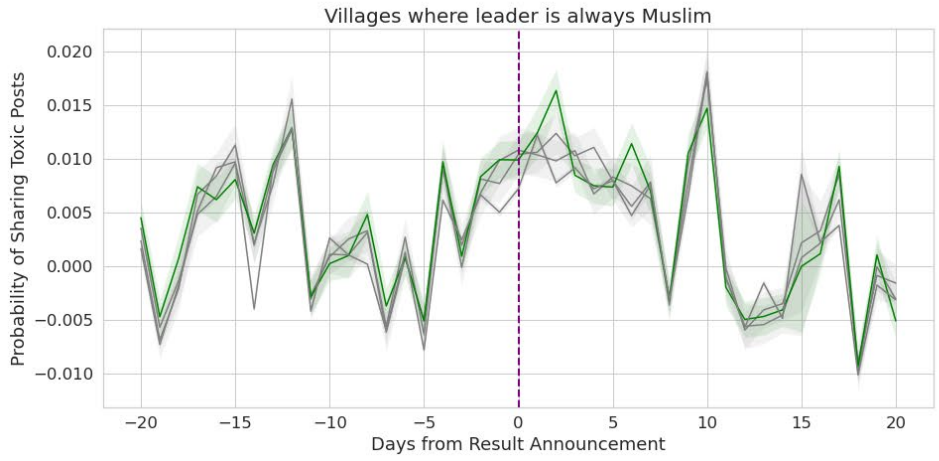
Name Classification

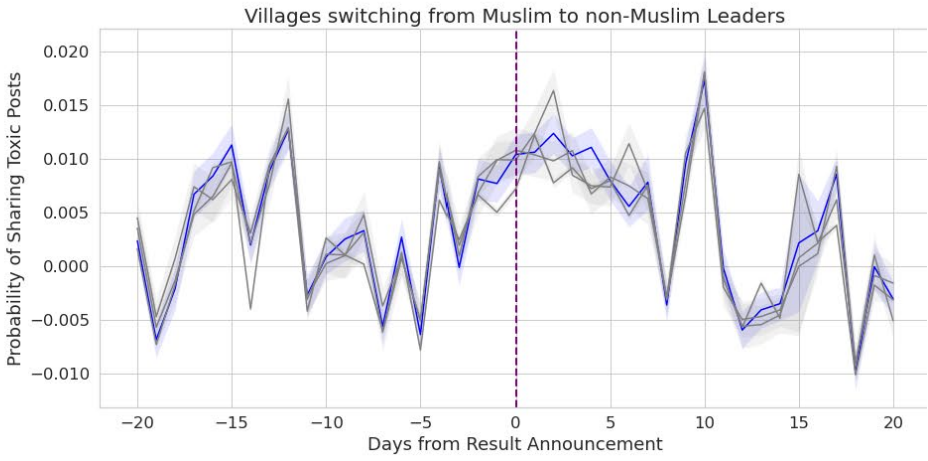
- Classified candidate religion using names
- Trained Neural network model on set of manually annotated names

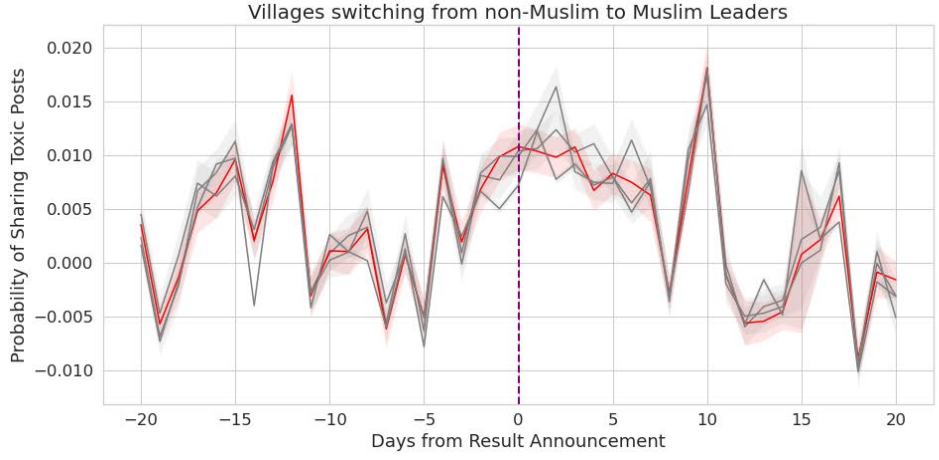
Year	Muslim Winner	Muslim Runner Up
2021	9.65	11.51
	(0.29)	(0.32)
2015	10.21	–
	(0.30)	–

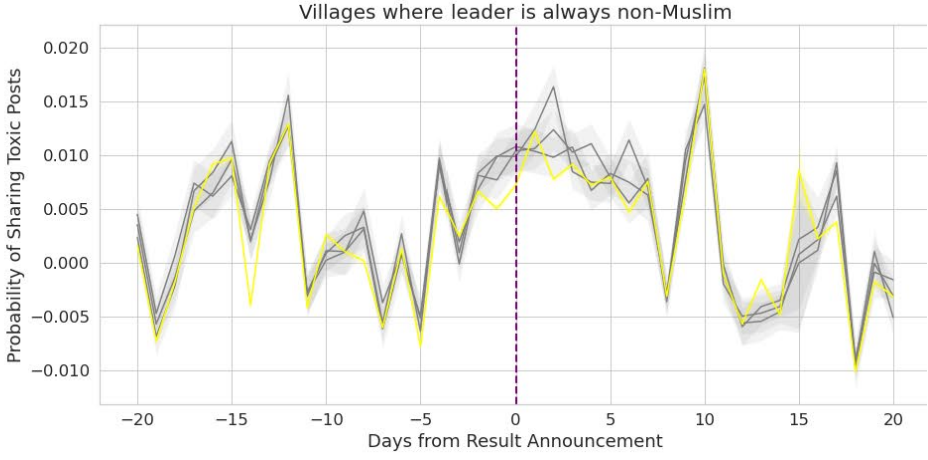
Muslim candidates as percentage share of all candidates

Engagement with Toxic Speech by Leader Identity









Empirical Strategy

Close Elections

$$y_{vt} = \beta \cdot 1(VM_v^{\text{muslim}} \geq 0) + f(VM_v^{\text{muslim}}) + \gamma X_{vt} + \delta_t + \varepsilon_{vt}$$

y_{vt}: average toxicity score of shared posts, in village v, on date t

VM_v: vote margin of Muslim candidates $\in [-l, h]$

X_{vt}: user attributes in village v at time t

δ_t: date fixed effects

Identifying Assumptions

Coefficient of Interest: β

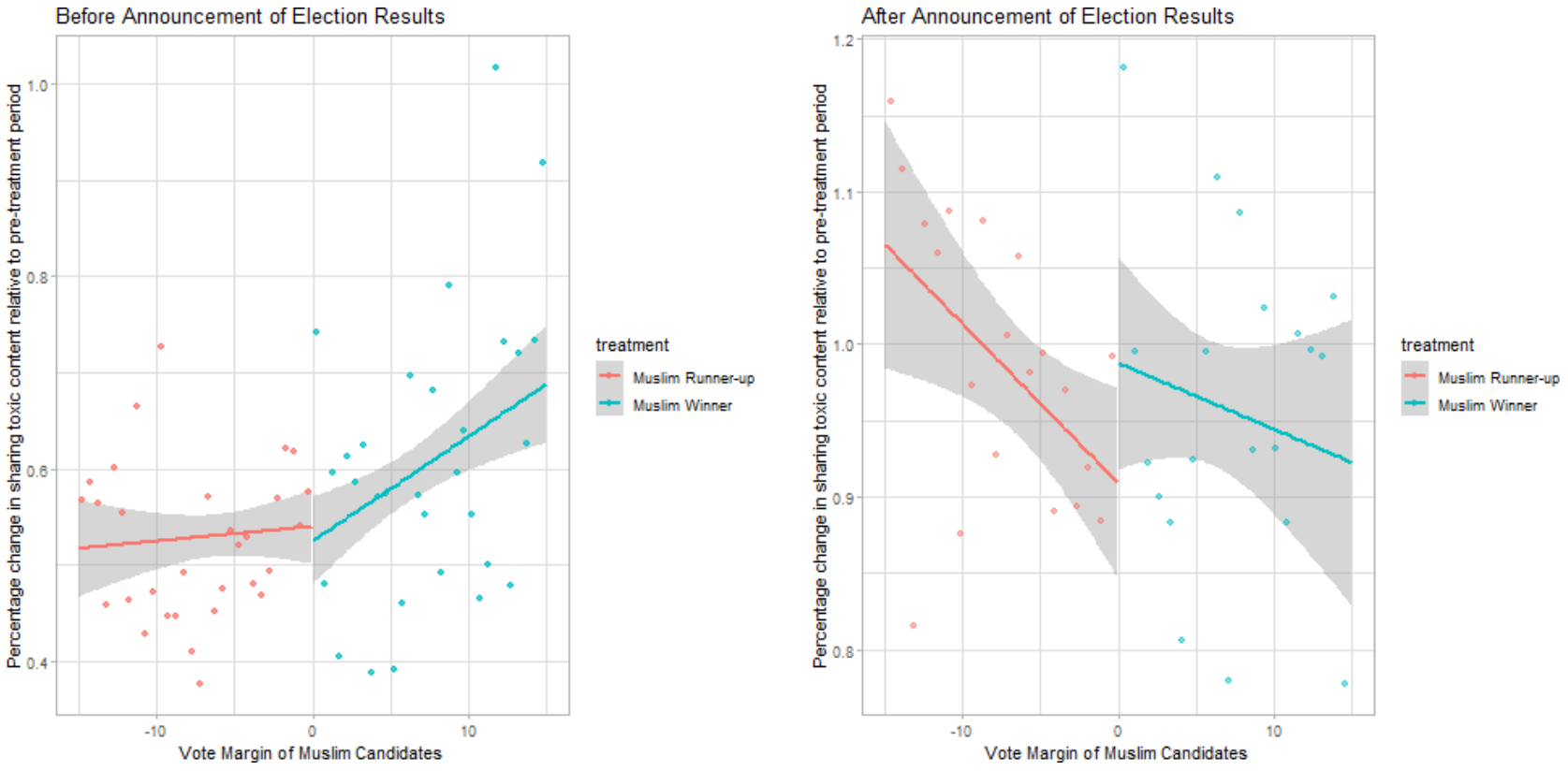
Potential Outcome functions are continuous at the cutoff (Hahn et al., 2001)

$$\mathrm{E}[y_{vt}(d)|VM_v^{\text{muslim}} = z] \text{ is continuous at 0 for } d = 0, 1$$

Recall that,

$$d = 1(VM_v^{\text{muslim}} \geq 0) = \begin{array}{l} 0 \text{ if Muslim candidate loses,} \\ 1 \text{ if Muslim cadidate wins.} \end{array}$$

Results



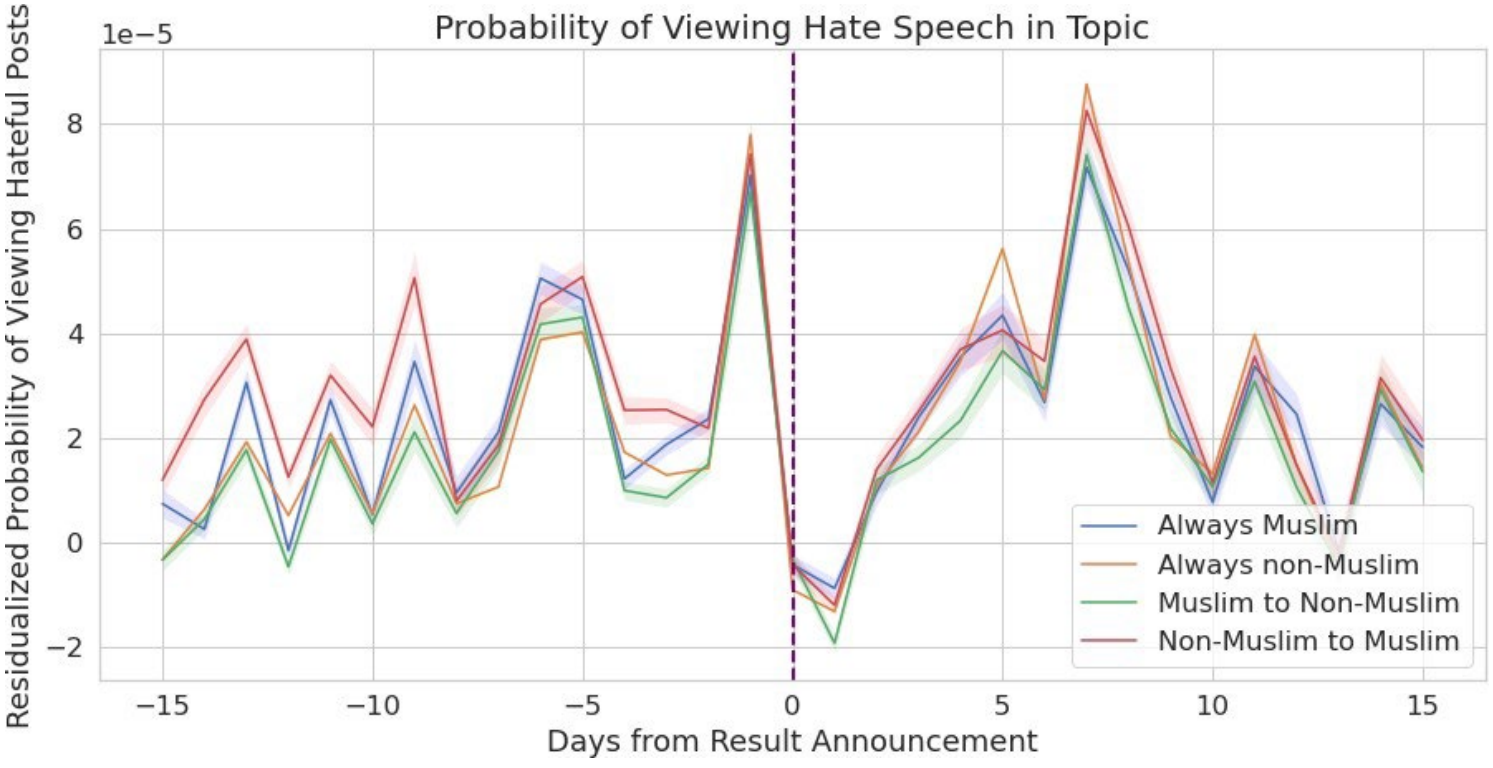
Percentage change in toxicity of shared posts in treated (Muslim winner) and control (Muslim runner-up) villages

	Pre-Announcement	Post-Announcement
Treatment	-0.06	0.198
(Muslim winner)	(0.12)	(0.202)
95% Confidence Interval		
Bias Corrected	[-0.301, 0.171]	[-0.145, 0.648]
Robust	[-0.343, 0.213]	[-0.21, 0.712]
Control Mean	0.517	0.867

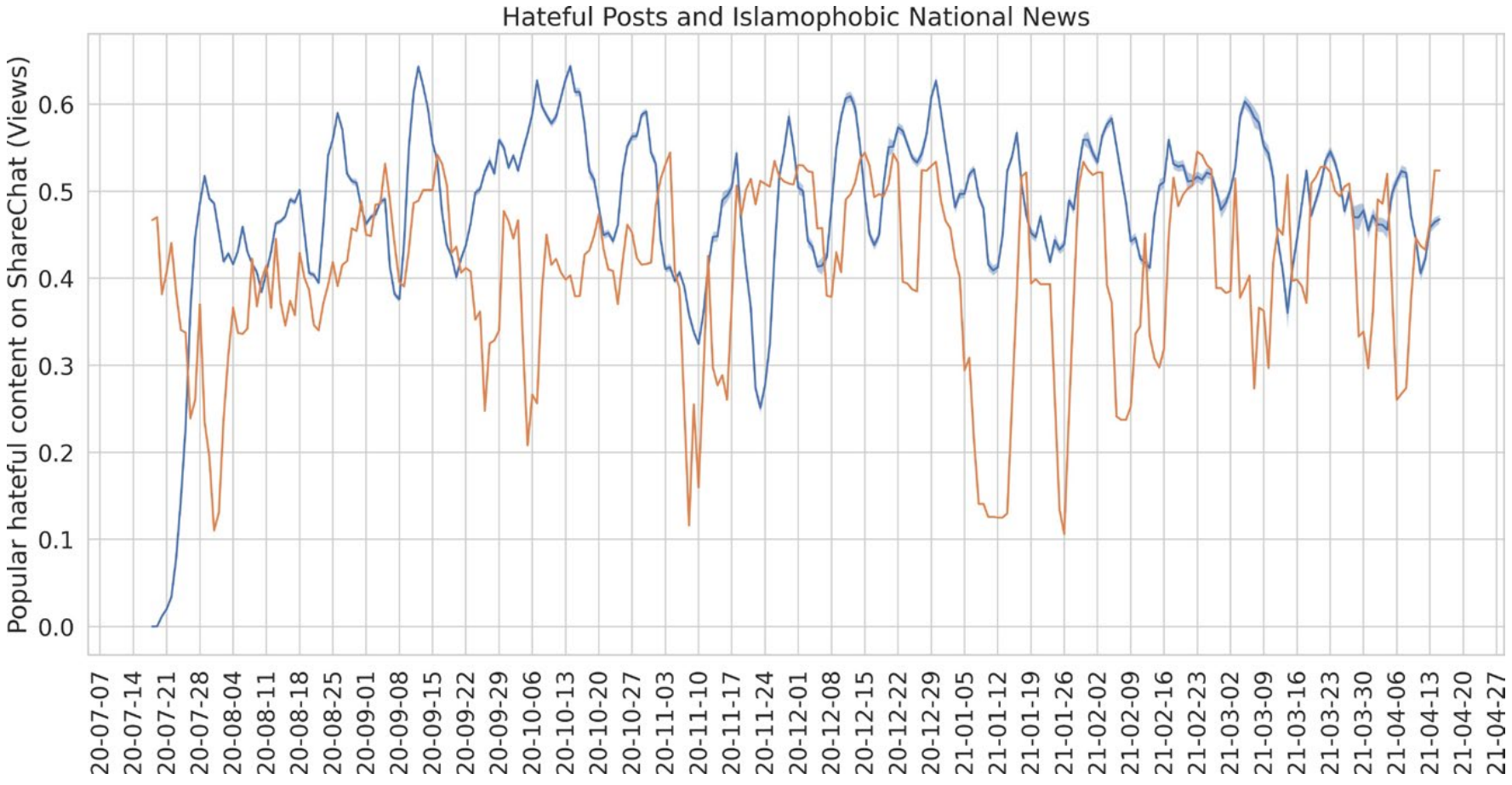
Local linear regression results on a random sub-sample of salient elections

Discussion

Correlated Patterns of Exposure

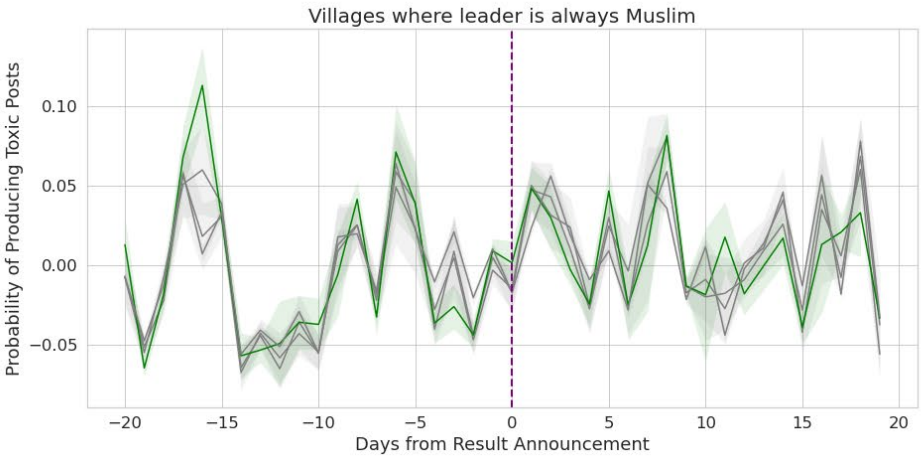


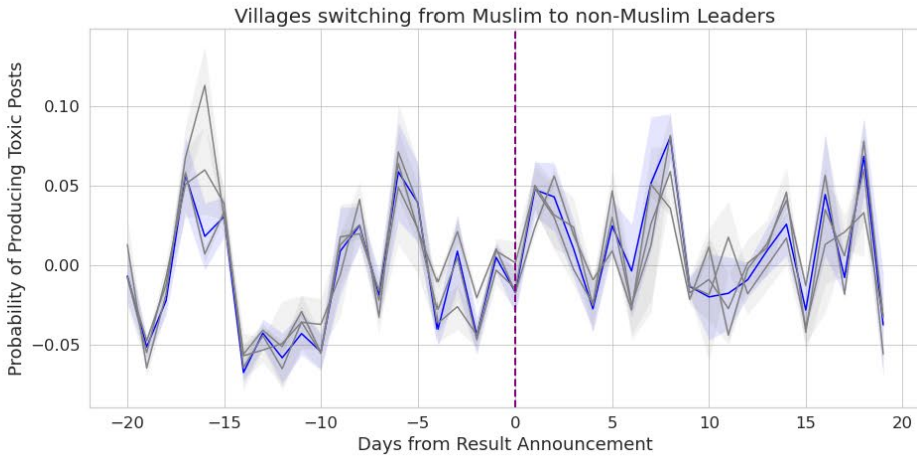
National, not Local

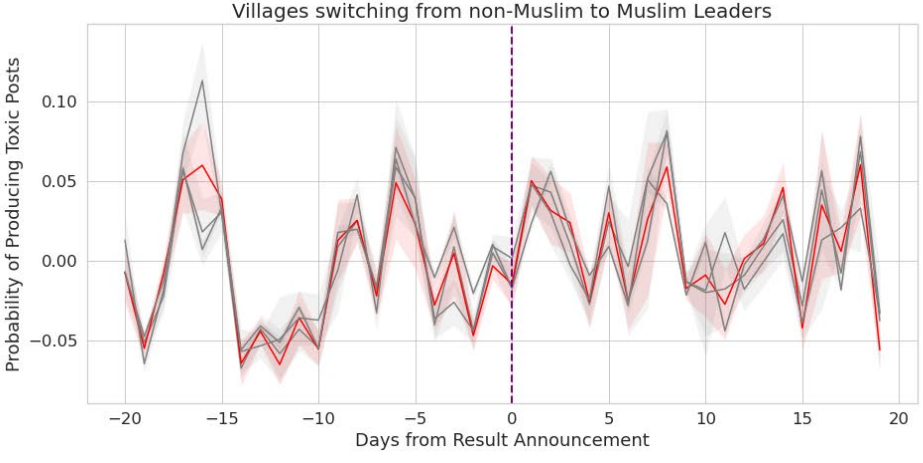


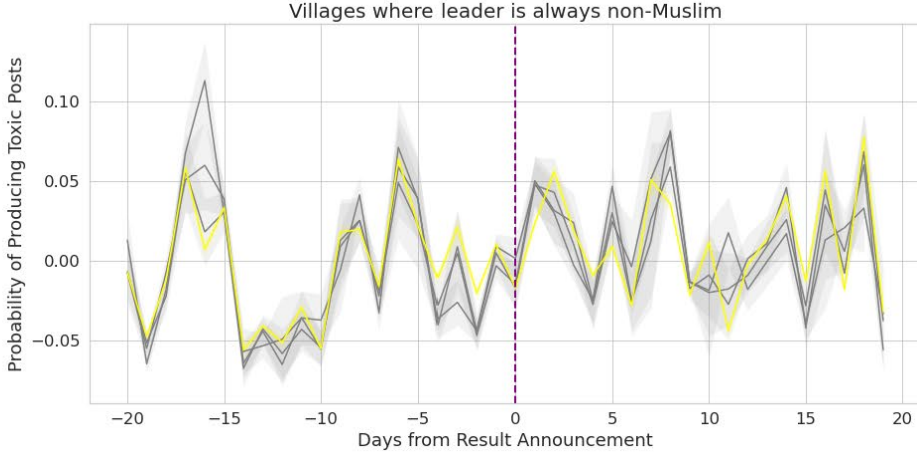
Trends in exposure to hateful content on ShareChat (blue) and engagement with OpIndia posts on Twitter (orange)

Production


Villages where leader is always Muslim
Probability of Producing Toxic Posts
0.10
0.05
0.00
−0.05
−20
−15
−10
−5
0
5
10
15
20
Days from Result Announcement


Villages switching from Muslim to non-Muslim Leaders
Probability of Producing Toxic Posts
0.10
0.05
0.00
−0.05
−20
−15
−10
−5
0
5
10
15
20
Days from Result Announcement

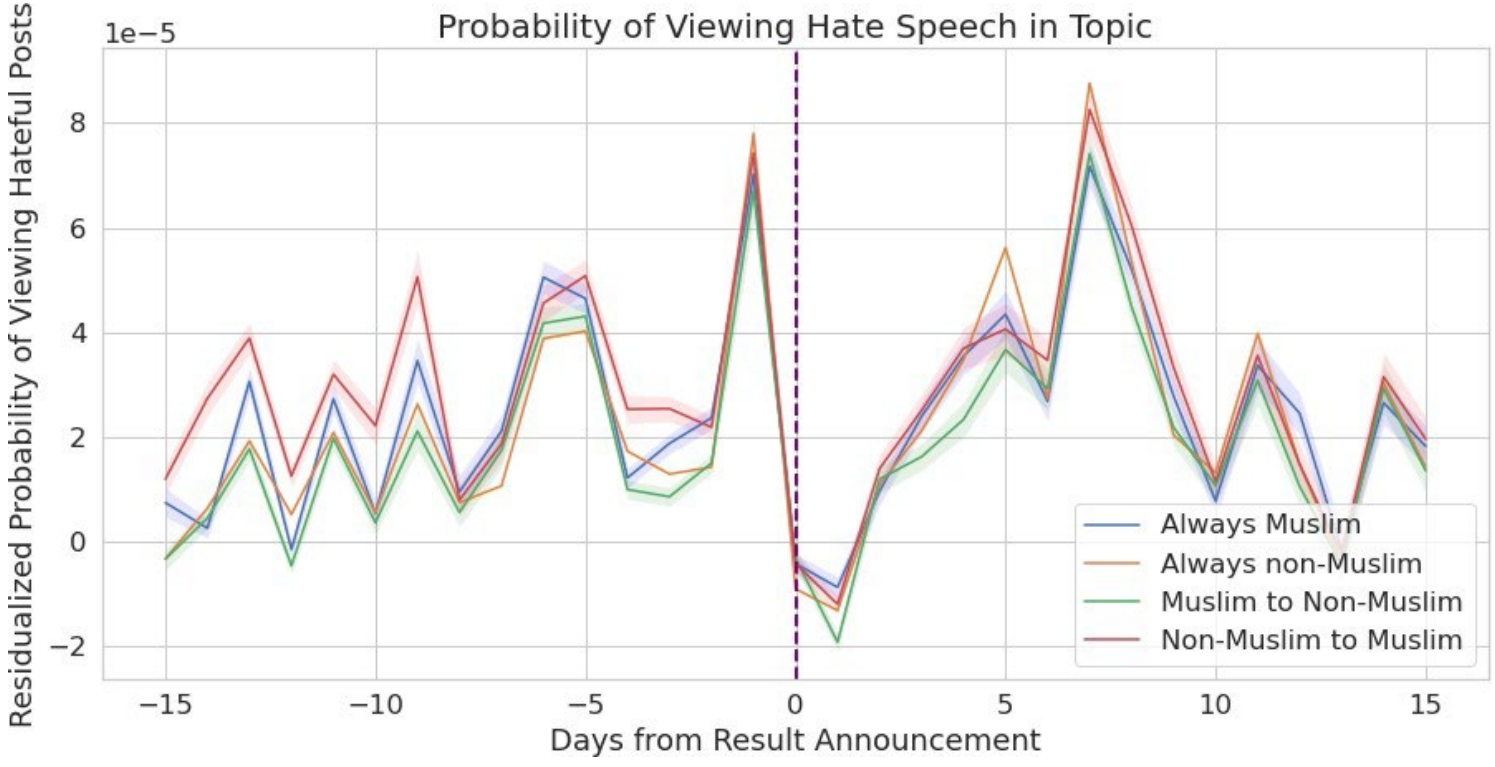

Villages switching from non-Muslim to Muslim Leaders
Probability of Producing Toxic Posts
0.10
0.05
0.00
−0.05
−20
−15
−10
−5
0
5
10
15
20
Days from Result Announcement


Villages where leader is always non-Muslim
Probability of Producing Toxic Posts
0.10
0.05
0.00
−0.05
−20
−15
−10
−5
0
5
10
15
20
Days from Result Announcement

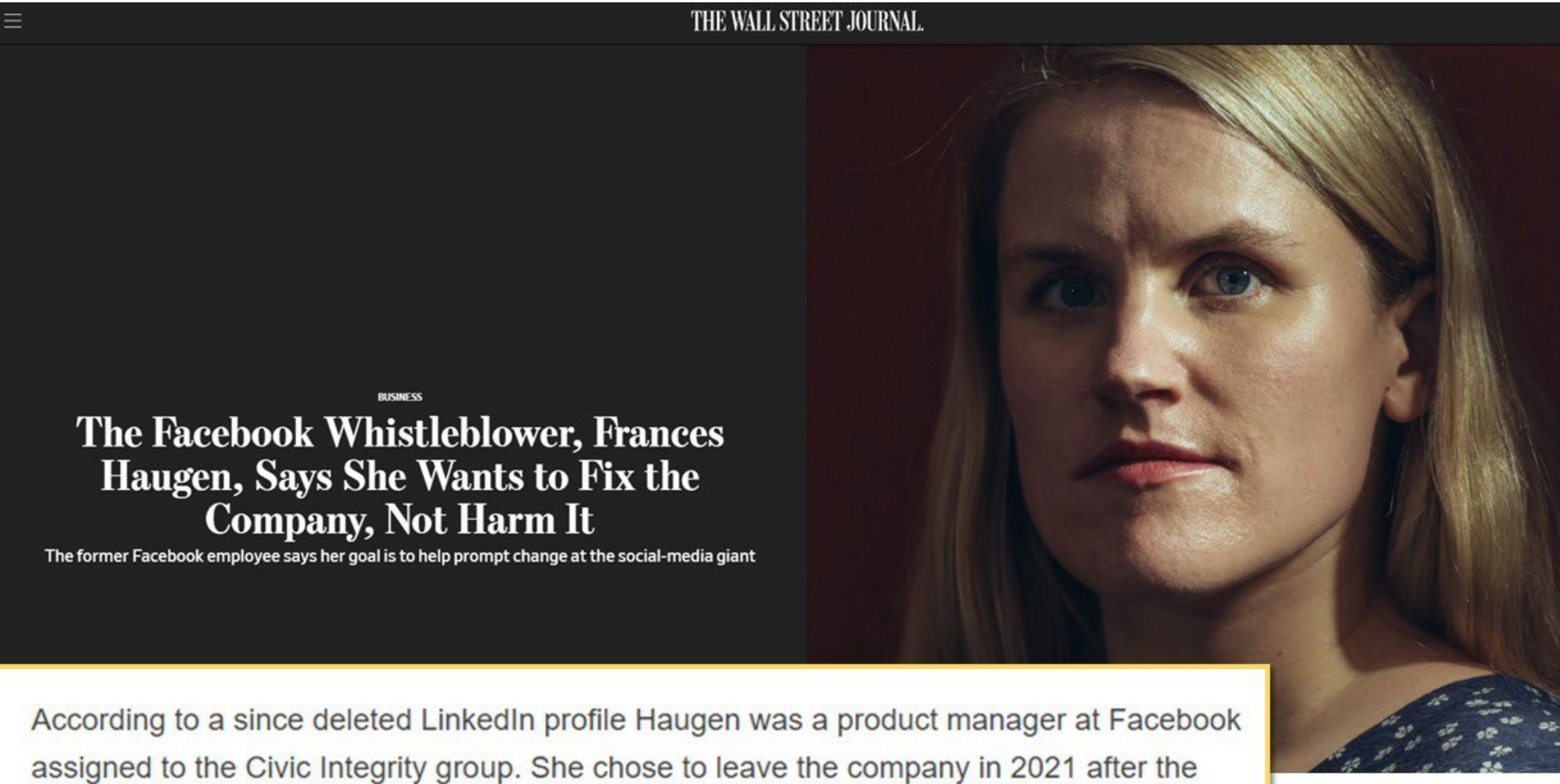
Hate Speech: Engagement on ShareChat

What *causes* engagement with hateful content on social media?

- Economic Shocks
- Political Shocks
- Algorithms

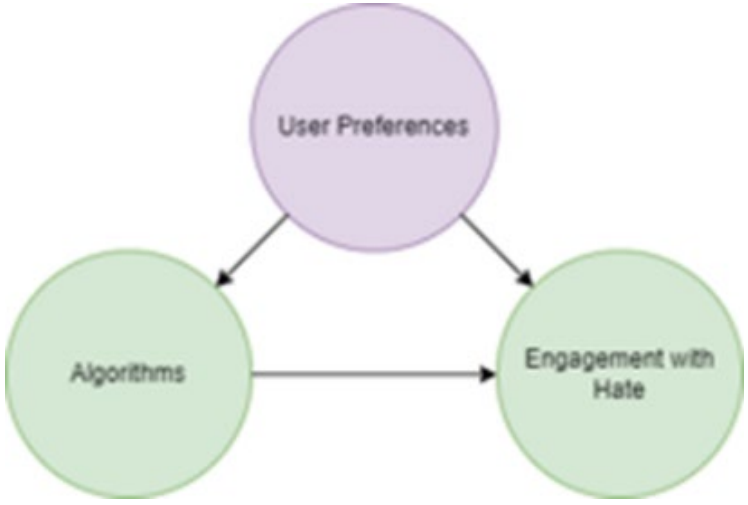


Regulating Platform Recommender Systems



According to a since deleted LinkedIn profile Haugen was a product manager at Facebook assigned to the Civic Integrity group. She chose to leave the company in 2021 after the dissolving of the group. She said she didn't "trust that they're willing to invest what actually needs to be invested to keep Facebook from being dangerous."

Causal Effects of Algorithmic Recommender Systems



- Users spending increasing amount of time on social media
  - Increased ad consumption
  - Increased consumer surplus
- Habit formation and digital addiction
- Network spillovers
- Increased political polarization
  - Ambiguous effect on consumer surplus

Experimentation with Algorithms

Candidate Generator	Creates large set of posts to be ranked 10,000 candidate posts per day
Ranker	Picks top 100 posts according to CG scores Scores to rank posts using more information

Conclusion

- Factors Driving Hate Speech
  - Election cycles
  - National political conditions
- No evidence of response to leader's identity
- Future and Present work
  - Algorithms!

References I

Acemoglu, D., Egorov, G., and Sonin, K. (2013). A political theory of populism. *The Quarterly Journal of Economics*, 128(2):771–805.

Ajzenman, N., Cavalcanti, T., and Da Mata, D. (2020). More than words: Leaders' speech and risky behavior during a pandemic. *Available at SSRN 3582908*.

Arun, C. (2019). On whatsapp, rumours, lynchings, and the indian government. *Economic & Political Weekly*, 54(6).

Benabou, R. and Tirole, J. (2011). Laws and norms. Technical report, National Bureau of Economic Research.

Bettinger, E. P. and Long, B. T. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, 95(2):152–157.

References II

Bhalotra, S., Clots-Figueras, I., Iyer, L., and Vecci, J. (2021). Leader identity and coordination. *The Review of Economics and Statistics*, pages 1–50.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2020). Cross-country trends in affective polarization. Technical report, National Bureau of Economic Research.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020a). From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–48.

Bursztyn, L., Haaland, I. K., Rao, A., and Roth, C. P. (2020b). Disguising prejudice: Popular rationales as excuses for intolerant expression. Technical report, National Bureau of Economic Research.

References III

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Suryawanshi, S., Jose, N., Sherly, E., and McCrae, J. P. (2020). Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

De, A., Elangovan, V., Maurya, K. K., and Desarkar, M. S. (2021). Coarse and fine-grained hostility detection in hindi posts using fine tuned multilingual embeddings. In Chakraborty, T., Shu, K., Bernard, H. R., Liu, H., and Akhtar, M. S., editors, *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 201–212, Cham. Springer International Publishing.

Enikolopov, R., Petrova, M., and Zhuravskaya, E. (2011). Media and political persuasion: Evidence from russia. *American Economic Review*, 101(7):3253–85.

References IV

Garimella, K. and Eckles, D. (2020). Images and misinformation in political groups: Evidence from whatsapp in india. *arXiv preprint arXiv:2005.09784*.

Gentzkow, M., Shapiro, J., Taddy, M., et al. (2016). Measuring polarization in high-dimensional data: Method and application to congressional speech. Technical report.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, 143:73–88.

References V

Jaffrelot, C. and Kalaiyarasan, A. (2019). The political economy of the jat agitation for other backward class status. *Economic and Political Weekly*, 54(7):29–37.

Kuziemko, I. and Washington, E. (2018). Why did the democrats lose the south? bringing new data to an old debate. *American Economic Review*, 108(10):2830–67.

Meyersson, E. (2014). Islamic rule and the empowerment of the poor and pious. *Econometrica*, 82(1):229–269.

Mitra, A. and Ray, D. (2014). Implications of an economic theory of conflict: Hindu-muslim violence in india. *Journal of Political Economy*, 122(4):719–765.

Müller, K. and Schwarz, C. (2019). Fanning the flames of hate: Social media and hate crime. *Available at SSRN 3082972*.