Hate Thy Neighbor: Effects of News Localization on Political Polarization in a Large-Scale Experiment

Aarushi Kalra[*]

November 25, 2024

Abstract

This paper examines the efficacy of non-invasive interventions in mitigating engagement with politically divisive content on social media platforms. Leveraging an individually randomized experiment with over 50 million users in collaboration with a major social media platform in India, I introduce "viewpoint-blind" content, nudging users with politically neutral, localized news stories, as an alternative to contentious content moderation practices and biased mainstream media sources. The experiment identifies the causal effect of these nudges on user engagement patterns, particularly focusing on interactions with harmful anti-minority content. Results indicate that increasing the visibility of neutral content significantly reduces engagement with divisive narratives, with a 3% decrease in interactions with polarizing content. This study offers policy-relevant insights for addressing online misinformation while balancing concerns over regulations infringing upon free speech and discouraging overall platform usage. These findings have important implications for platform governance and the design of algorithms that shape information consumption in digital spaces.

Keywords: Digital Economies, AI Policy, Media Bias, Local News

[*] Department of Economics, Brown University, Providence, RI, 02906 (email: aarushi_kalra@brown.edu). I am grateful to Andrew Foster, Brian Knight, Peter Hull, and Daniel Björkegren for their continued guidance and support. This project has greatly benefitted from helpful discussions with Pedro Dal Bo, Ro'ee Levy, Jon Roth, Stelios Michalopoulos and seminar participants at Brown University. I thank Ahad Bashir and Neha Narayan for excellent research assistance. All remaining errors are my own.

1 Introduction

The proliferation of harmful discourse on social media platforms poses significant societal challenges (Duggan, 2014). These narratives in the digital space have also translated into physical violence against vulnerable communities (Müller and Schwarz, 2021, 2023). While government regulation can help in addressing these issues, content moderation policies often face criticism when they are seen as infringements on free speech. In the United States, conservative groups primarily voice these concerns, viewing such interventions as targeted censorship. However, in India, progressive groups argue that their civil rights to freedom of expression have been compromised. As a result, content moderation policies face criticisms from both the ends of the political spectrum (Kominers and Shapiro, 2024).

This tension has spurred research into "viewpoint-blind" approaches for moderating content without resorting to post removal or "deplatforming," a particular individual or ideology.[1] While previous studies, such as Katsaros et al. (2022), have explored interventions like prompts encouraging users to reconsider reposting of potentially harmful content, there is little evidence on the effectiveness of possible viewpoint-blind policy instruments.

In this paper, I use an individually randomized experiment with 50 million users conducted by a major Indian content generation platform analogous to TikTok. I partner with this platform (henceforth, SM) with nearly 200 million active monthly users across a dozen regional languages to study the imoact of view-point-blind content moderation policies. Recent literature has linked social media discourse in India to pressing concerns about anti-minority hate crimes and mob violence (Banaji et al., 2019). This underscores the urgency of evaluating interventions that can mitigate engagement with such content in this setting. Still, despite India being the second-largest market for social media globally, it remains understudied.

Nudges, conceived in this paper as insertion of local news posts in content feeds, while often considered powerful and low-cost tools for changing economic outcomes, may have unintended consequences (Thaler and Sunstein, 2021; Hume et al., 2024). This paper contributes to the growing body of research on viewpoint-blind interventions by examining a novel approach: promoting politically neutral, localized news content into users' feeds. I hypothesize that this method can reduce engagement with divisive content without explicitly targeting specific viewpoints, potentially avoiding the pitfalls associated with more direct forms of targeting.

The choice of local news as an intervention tool is motivated by several factors. First,

[1] The idea of viewpoint-blind regulation was first introduced by Supreme Court Justice Clarence Thomas. See `https://theweek.com/articles/975842/clarence-thomas-enigma-social-media`.

local news posts are orthogonal to the content SM users are accustomed to seeing as these posts are unlikely to be recommended by the feed-ranking algorithms. Locally relevant content provides utility to a specific set of users but is unlikely to be promoted by recommendation algorithms programmed to maximize overall content engagement. This is because local content doesn't capture the same returns to scale as national news. Second, in my experimental setting, the platform did not employ user location to generate customized recommendations at the time of this intervention, further reducing the likelihood of local news being algorithmically promoted. Here, local news is generated by local journalists reporting on job openings, traffic alerts, and local amenities. Third, these posts are not hateful by construction, ensuring that the intervention remains viewpoint-blind and politically neutral.

The experimental design comprises two treatment arms and a control group, containing approximately 1.5 million users consuming content in the Hindi language. Treatment was randomly assigned as the user-level. The treatment arms differ in the level of effort required for users to access local news content: the "Trending Feed" arm displays local news on the platform's landing page, while the "Bucket Feed" arm provides access through a new tab on the home screen. The control group receives zero exposure to local news content during the intervention period. This design allows for causal identification of the impact of increased exposure to neutral content on user engagement with "toxic" anti-minority material, as well as exploration of potential underlying mechanisms. I define toxicity according to a Google-developed algorithm tailored to measure harm a post can cause to vulnerable groups.[2]

The main outcome of interest is engagement with harmful content. This presents a significant measurement challenge, as the technology for multi-lingual and context-sensitive hate speech detection is still evolving (Roy et al., 2021). To address this challenge, I first translate close to six million posts to English using the Google Translate API. I then employ a "toxicity" detection algorithm from Perspective API, created by Google and Jigsaw. The Perspective algorithm provides the current best machine learning solution for toxicity detection, relying on training data containing millions of toxic and non-toxic comments, marked by human raters (Jiménez Durán, 2022). I observe engagement (the action of sharing a post off SM to other platforms like WhatsApp) with user-generated content that is assigned toxicity scores in the administrative data from SM in 2021.

I focus my attention on anti-minority toxic content, by considering toxicity of a subset of

[2] In particular, I examine the intersection of toxic and political posts that are shown to verbally attack India's Muslim minority. Toxicity, as defined by Google's Perspective API, measures a post's hatefulness or potential harm. This API, used by organizations like the New York Times and in academic research (Beknazar-Yuzbashev et al., 2022), defines a toxic comment as a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion." The algorithm relies on millions of comments labeled by human raters to learn patterns of toxic behavior, using a semi-supervised learning approach. For more information, see `https://perspectiveapi.com/how-it-works/`.

posts that relate to Political, Religious, Cultural or News content genres.[3] I show that toxic content in these genres is largely targeted at Muslims, and such posts carrying misinformation have been linked with anti-Muslim hate crimes in India (Arun, 2019). By examining the effects of this local news intervention on engagement with toxic content, this study aims to contribute to our understanding of viewpoint-blind content moderation strategies and their potential to mitigate harmful discourse on social media platforms.

I find a 3% reduction in engagement with toxic content among users assigned to the trending feed intervention. I show that this decrease is not driven by differences in platform usage in the treatment groups. My first stage regression estimates of treatment on exposure to local news and toxic content show that treated users saw about 5 more local news posts, compared to users in the control group who saw none. This translates into a 2.9% reduction in toxicity of content exposure. On the other hand, although bucket feed treatment reduced exposure to toxic content by 1.4%, there were no significant changes in toxic sharing among users randomly assigned to this treatment arm.

This paper makes a methodological contribution by devising a new test for checking if user behavior is "mechanical." Users are said to behave mechanically when they share a constant proportion of a particular type of content, with and without the intervention. The rate of sharing among treated users cannot be measured using a log specification, as these estimates depend on the unit of measurement (Thakral and Tô, 2023). This makes estimates unreliable in the presence of a large number of zeroes in the data (Chen and Roth, 2024). In addition to proving the validity of this statistical test with a wide variety of applications, I show that the intervention induced behavioral changes, and that the treatment effect is not mechanical. Instead, experimentally exposing users to neutral content induces behavioral changes on the platform.

This paper contributes to three strands of the literature. First, this paper relates to a rich literature on media bias, and its impact on human behavior (DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2011; Chiang and Knight, 2011; Martin and Yurukoglu, 2017). In particular, my work closely relates to the literature on local news accessed through traditional media sources and political attitudes (Snyder Jr and Strömberg, 2010; Angelucci et al., 2024). However, the supply of such information is itself endogenously determined by a personalization algorithm in the case of social media platforms. I expand the scope of this literature by experimentally crowding out toxic content that may have been recommended by the personalization algorithm, in accordance with prior usage.

Second, this paper contributes to a large literature on the economics of social media (Aridor et al., 2022). The literature finds that social media negatively impacts mental

[3] These genres are determined automatically by the platform.

health outcomes as well as user welfare (Allcott and Gentzkow, 2017; Allcott et al., 2022; Braghieri et al., 2022). Furthermore, the relationship between violence against vulnerable communities and social media has now been systematically documented (Müller and Schwarz, 2021, 2023). On the other hand, a related strand of the literature also shows the positive effects of social media by exposing users to alternative view-points (Gentzkow and Shapiro, 2011; Carney, 2022). I study an important, yet understudied, population in a context where personalization algorithms have been introduced recently on social media platforms, and regulations are scant. The case of India is also pertinent due to an urgent need to moderate content that has been linked to offline violence against minority communities.

Finally, this paper is directly linked to a burgeoning literature on news consumption through social media (Levy, 2021; González-Bailón et al., 2023). To the best of my knowledge, this paper is the first to use local news as an intervention to reduce users' engagement with polarizing content. Moreover, the large sample sizes and multiple treatments in my study not only enable me to precisely estimate the direct effects of the intervention, but also to measure the non-linearities in the effect of reduced exposure to toxic content. Consistent with these papers, I find that diversifying news sources improves political behaviors.

However, my work complements this literature by showing the effectiveness of a policy instrument that reduces engagement with toxic content, without affecting platform usage. Crucially, my study is free from experimenter demand effects and selection. This is because users did not know that they were part of an experiment, as they consent to be studied for market research, at the time of account creation. This significantly contributes to our understanding of socially undesirable behaviors, as well as policies to check them, outside of lab-like settings.

The remaining paper is organized as follows. Section 2 provides background details of the context and Section 3 provides details of the administrative as well as experimental data employed in this study. Section 4 delineates the design of the experiment, and provides descriptive statistics that motivate an interrogation of mechanisms with a structural model. Section 5 provides the empirical framework to verify if treated users behave mechanically, and to trace the shape of the treatment effects. Section 6 presents the main results from an empirical analysis of the experiment. Finally, Section 7 concludes with a discussion of the implications of the results for technology policy.

2 Background

I study the effect of inserting local news content in user feeds on engagement with toxic content on an Indian social media platform. The context of my study is important because a major share of the Indian population is on social media, discourse on these platforms has serious implications for society. However, regulations are few, and weakly enforced.

2.1 Social Media and Indian Society

With over 500 million social media users, India is the second largest market for tech platforms in the world (Statista, 2023). Social media usage has grave consequences for the political environment in this setting, where social media posts are known to have provoked instances of violence, in the form of mob lynching, riots, and hate crimes (Banaji et al., 2019).

The mob lynching of Mohammad Akhlaq in 2015, just outside of the National Capital Region of Delhi, highlighted the role that platforms like WhatsApp play in spreading misinformation and exacerbating hate (Arun, 2019). This unfortunate incident is by no means an isolated one. Social media platforms like WhatsApp, Facebook, and Twitter face an unprecedented challenge of moderating content in this large market. This is difficult because these companies operate in a vastly different context, where hate speech on social media propagates in a very typical way. Furthermore, behavioral responses to various interventions also may be different among this population.

Context-driven content moderation is a difficult challenge also because the production of hate in the Indian context is very often linked with institutions that enable these platforms to carry out their businesses. Therefore, social media platforms may in fact be biased against opposition parties and pressure groups. This was seen when Twitter decided to suspend various accounts linked with the Farmer's Movement during massive protests against three controversial farm bills that were passed by the Indian Parliament (Dash et al., 2022). Similarly, the Wall Street Journal has alleged that Facebook India's Public Policy Head selectively shielded offensive posts of leaders of the ruling Bharatiya Janata Party (BJP), which has been described as Prime Minister Modi's Hindu Nationalist Party. According to WSJ, BJP leader T. Raja Singh, has said in Facebook posts that Rohingya Muslim immigrants should be shot, called them Muslims traitors and threatened to raze mosques to the ground (Purnell and Roy, 2020).

However, it is very difficult to detect political bias in platform content moderation policies when various groups across the political spectrum have felt targeted by them (Kominers and Shapiro, 2024). India's Information Technology Minister accused Facebook of bias against supporters of right-wing viewpoints when the platform banned the said BJP leader accused by the WSJ (for propagating anti-minority hate speech in digital commons).[4] Platforms do not want to moderate political content with given slant for being accused of leaning towards

[4] See `https://www.wsj.com/articles/facebook-faces-hate-speech-grilling-by-indian-lawmakers-after-journal-article-11597747734`

one view-point or another. This accentuates the need for view-point-blind policy instruments or products.

2.2 The Platform

I partner with SM, a rapidly growing social media platform in India, to understand the effects of view-point-blind policies like injection of local news in user feeds. SM's app features resemble those of TikTok, and the platform made massive gains in market share when TikTok was banned in India.[5]

I study how the nature of online interactions changes with the intervention in SM's rich online social network. SM is one of the most popular platforms in the country with about 200 million active monthly users, who create and share content in around a dozen regional languages on the platform. Content based social networks, such as SM, are centered around topics like Romance and Relationships, Politics, Religion, and Greetings. The interactions on SM are mostly conducted through the 'trending' feed (Figure B.1). User feeds are algorithmically customized according to preferences revealed via past engagement with different types of posts.

SM is comparable to TikTok in its content generation and engagement features, and it attracts a large portion of the urban and rural poor populations in India. This makes such analysis especially important as little is known about political behavior of this demographic in India or about the users of this massive platform. Examples of popular political content (translated to English) are provided in the word cloud in Figure B.2. It is worth noting that highest weight (as computed by the highest number of text appearances in image posts) among posts with higher toxicity scores is given to the text *Ram*–which is associated with a deliberately polarizing slogan of the Hindu Nationalist organizations in India.[6]

3 Data

I bring administrative data from SM to understand how the intervention decreases engagement with harmful content on social media.

[5] TikTok was banned in India in 2020 due to rising geo-political tensions with China. See `https://www.nytimes.com/2024/03/22/business/tiktok-india-ban.html`.

[6] [tr.] Hail Lord Ram: these three words have also been referred to as the most polarizing words in India `https://foreignpolicy.com/2020/02/13/jai-shri-ram-india-hindi/`

3.1 Administrative Platform Data

I use administrative data from the platform to construct my main outcome variables. These include viewership of and engagement with political and toxic content at the user level. Engagement is measured using observed sharing behavior when a user decides to share the post with their social network on other platforms like WhatsApp. I limit the analysis to posts related to News and Politics.

Platform data include user characteristics, like their location, gender, and language. Administrative data provides information on each post that is viewed or engaged with by any given user. This allows me to trace the posts a user was exposed to, and whether they chose to engage with it. I can, therefore, condition engagement statistics on viewership, which is a feature absent from most observational studies that use social media data (Hosseinmardi et al., 2020). In this way, I show that the intervention not only changes what the users engaged with, but also changed what they were exposed to because of the personalized recommendations that were generated for them by this algorithm.

Content engagement on SM does not depend on social networks in terms of a user's "friends," followers or accounts they follow. This is because the personalization algorithm almost entirely depends on user interactions with different types of content in the past. Such platforms are seen as content-based networks as personal and horizontal interactions on the platform are not important in making content recommendations. This means that there were no spillovers of the intervention from the treated to the control as treatment assignment was done at the individual level. I provide evidence to support this in the following section.

3.2 Toxicity Classification

The administrative data provides user-post level data on viewership and engagement. To measure the main outcome variable, i.e. toxicity of posts engaged with, I further process the text from these posts in the admin data to classify them as being harmful.

Multilingual text classification is a key part of my research because I study the impact of my intervention on hatefulness of platform discourse and the content data is in Hindi. I develop a machine learning pipeline to first translate six million SM posts to English using the Google Translate API. I then use Perspective API to identify toxicity in the translated text. Toxic content is defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion[7]."

Perspective API, from Jigsaw and Google, provides the current best machine learning solution for toxicity detection, as it relies on training data from millions of comments from

[7] For more information, see `https://perspectiveapi.com/how-it-works/`

different publishers that are annotated by ten human rates on a scale of "very toxic" to "very healthy" contributions. The algorithm is based on Transformer based technologies like BERT, making the classification sensitive to the respective contexts (Hosseini et al., 2017). These models score a phrase "based on the perceived [negative] impact the text may have in a conversation."

I use the continuous toxicity scores to construct a measure of hatefulness towards Muslims for my analysis because I only consider political posts. Leading media houses like New York Times, and social media platforms like Reddit, have adopted this technology.[8] Perspective's Machine Learning models are, therefore, being widely adopted to identify and filter out abusive comments on various platforms.

3.3 Local News

Starting in 2020, SM made a concerted effort to hire local stringers to report on local news stories on local development issues like roads, COVID contamination zones, and employment opportunities. Therefore, in my experimental setting, local news is generated by local journalists reporting on job openings, traffic alerts, and local amenities, for instance.

I describe the contents of local news posts using an LDA topic model in Figure B.3 (Ash and Hansen, 2023). These human-curated posts used in this intervention are not hateful by construction. I provide examples of local news posts, and contrast the toxicity scores on these posts against other political posts in Figure B.4.

Since locally relevant content provides utility to a small set of users, it is very unlikely to be promoted by algorithmic recommender systems that are programmed to maximize content engagement. Such content does not capture the same returns to scale as national news. Additionally, these posts are not recommended by the algorithm because SM did not employ user location to generate customized recommendations at the time of this intervention.

4 Experimental Design

I lay out the design of the experiment below.

4.1 Treatment

The intervention randomly assigns users to one of two treatment arms, which differ in the time and effort spent to get local news on the platform. Treated users were either exposed to local news on the landing page of the platform (in the treatment arm called *Trending Feed*),

[8] See `https://perspectiveapi.com/case-studies/`.

or they saw a new tab on their home screen to access local news stories (in the treatment arm called *Bucket Feed*). Users in the control group were not exposed to any local news content during the intervention period, either through the trending feed or through a new tab.

The trending feed is highlighted in Figure B.1 which shows the user interface. The tabs to the left of the trending feed tab are typically specific to a topic. Users assigned to the Bucket Feed treatment arm see an additional tab to the left of the trending tab, on top of the landing page.

Figure 2 shows that an average treated user, in either treatment arm, viewed about 5 local news posts on average while control users sat none. This shows that the first-stage is strong, but the intervention may not have been very salient. Figure B.5 shows the distribution of views on local news posts across the three groups.

4.2 Sample

Out of an experimental sample of over 55 million users, I analyze a sample of close to 1.4 million users. This is because I only retain users who use the platform in Hindi language to validate the text analysis, as this is the language I am most proficient in. Hindi-language users constitute less than 20% of the platform's user-base.

Further, I consider active users in the analysis or users who viewed at least 200 posts in the month of December 2020, which serves as the baseline period. Then, there are 210824 users in the control sample, 313944 in Bucket Feed, and 784127 in the trending feed sample, as shown in Table 2.

4.3 Randomization

I validate that users are randomly assigned treatment by showing that probability of treatment assignment is uncorrelated with observable user characteristics, like their city of residence, gender, and age. I construct balance tables by estimating the following regression equations in the constructed sample.

$$D_i^{bf} = \beta_0 + \sum_c \beta_c \mathbf{1}_i(characteristic = c) + \varepsilon_i$$

$$D_i^{tf} = \beta_0 + \sum_c \beta_c \mathbf{1}_i(characteristic = c) + \varepsilon_i$$

where, D_i^{bf} is a dummy variable that takes the value 1 when user i is assigned the Bucket Feed treatment, i.e. a treated user can access local news content in a separate tab. Similarly,

D_i^{tf} is an indicator for Trending Feed treatment, i.e. a treated user is exposed to local news content on the landing page. Validity in randomization requires that β_c are all individually and jointly insignificant.

Table 1 shows that I cannot reject the null hypothesis of joint insignificance of the beta coefficients for treatment assignment to either arms. Therefore, treatment assignment is uncorrelated with gender and age of the user. Since random treatment assignment was done at the user level, the average number of local news posts supplied is expected to be indistinguishable across treatment and control groups during the intervention period. This is because Table 1 shows that users' city of residence is not differentially correlated with treatment assignment.

4.4 Descriptive Statistics

Local news posts are created by journalists at the sub-district level, and Table 2 shows that on average, each location cluster was supplied with close to 10,000 local news posts over March 2021, which is the main intervention period I analyze. Prior to the intervention, in December 2020, the number of posts viewed by users across treatment and control units was 43 posts on average over the entire month. This table also shows that there is balance in total number of posts viewed at baseline.

The experiment helps answer the research question on how can engagement with toxic content be reduced without censoring content. I explore if such changes in user behavior are driven by differences in exposure to content types, and not by altering aggregate behavior on the platform, such as total number of posts viewed or the number of times users log onto the platform. Although differences in number of posts viewed are statistically significant in Figure 1, these differences are not economically significant. This is because the decline in total usage ranges from 2% to 3%, which is much smaller than the 27% reduction in total views in Kalra (2024) where I ran a much more drastic intervention on the same platform. Similarly, Figure B.6 shows that the maximum decrease in the number of times users log onto the platform is 0.4%, whereas this effect equals 6% in the companion paper.

This is important because prior research measuring the effect of exposure on online behavior impacts outcomes by changing the total number of posts viewed or the time spent on the platform (Guess et al., 2023; Beknazar-Yuzbashev et al., 2022). This intervention has the advantage that the intervention effects behavior only through the channel of displacing toxic posts without impacting overall platform engagement. In this way, the paper makes a crucial contribution to the existing literature.

4.5 Treatment Intensity

This intervention provides a rare opportunity to study a policy that does not alter over all platform usage, but reduces users exposure to toxic content. Therefore, the exposure to local news posts is interpreted as the relvant treatment intensity. Figure 2 indicates a strong first stage. I formally test this by estimating the following regression equation.

$$localViews_i = \beta_0 + \beta_1^{bf} D_i^{bf} + \beta_1^{tf} D_i^{tf} + \varepsilon_i \tag{1}$$

Here, $localViews_i$ corresponds to the number of local news posts viewed in one specification, and it is an indicator function taking value 1 if user i viewed any local news once every two days on average, in another specification. Crucially, D_i^{bf} and D_i^{tf} are indicator functions, taking value 1 if user i has been randomly assigned to the Bucket Feed or Trending Feed treatment arms respectively.

Table C.1 presents the estimated regression coefficients in (1) above. This shows that the first stage is statistically significant as treated users view more local news posts during the intervention period. Both Bucket Feed and Trending Feed users viewed approximately 5 more local news posts in one month. This table also shows that the probability of viewing any local news positive is strictly positive in both treatment arms. However, column (2) notes that this probability is significantly less than 1. For this reason, the main effects should be interpreted as the intent to treat.

4.6 Toxic Views and Shares

I am interested in the effect of the intervention on users' online behavior with respect to harmful content. That is, I analyze the effect of the intervention on the average toxicity of user feeds (or the average toxicity of their views) and the average toxicity of the posts they choose to share. That is,

$$ToxicView_i = \lambda_0 + \lambda_1^{bf} D_i^{bf} + \lambda_1^{tf} D_i^{tf} + \mu_i \tag{2}$$

$$ToxicShare_i = \pi_0 + \pi_1^{bf} D_i^{bf} + \pi_1^{tf} D_i^{tf} + \nu_i \tag{3}$$

where, $ToxView_i$ and $ToxShare_i$ represent average toxicity of user i's views and shares, respectively. As before, D_i's are dummy variables reflecting treatment assignment. Figure 3 and 4 show that treated users were less likely to view and share toxic posts, and the effect was larger for users in the Trending Feed. This is because the search costs for local news posts are higher for users assigned to the Bucket Feed arm. The intervention is successful in reducing engagement with toxic content by nudging users towards local news posts, away from the usual content feed where users find toxic content according to their preferences.

4.7 Mechanical Effects

It is conceivable that users learn social norms around political discourse from the content recommended to them by the algorithm. For example, if users view more toxic content on their feed, they may be influenced to believe that it is socially acceptable to say more toxic things against minority users.[9] It is also plausible that users share fewer toxic posts because they view fewer toxic posts to choose from.

If users reduce their toxic shares simply because they share a fixed proportion of toxic posts they view, they are said to behave "mechanically." Formally, under the constant effects assumption, a user is said to behave mechanically if she shares a constant proportion of the toxic content she views, irrespective of treatment status.

Proposition 1. *Consider a special case of the model so that sharing behavior does not respond to treatment assignment or the exposure to toxic content during the intervention period. Further the probability of sharing toxic content, for given levels of toxic exposure, is constant across all users, so that all users behave mechanically. Then the proportionate change in sharing toxic content is exactly equal to the proportionate change in toxic content viewed, i.e.*

$$\frac{\lambda_1^{bf}}{\lambda_0} = \frac{\pi_1^{bf}}{\pi_0} \quad \textit{and} \quad \frac{\lambda_1^{tf}}{\lambda_0} = \frac{\pi_1^{tf}}{\pi_0}$$

Proof. In Appendix A. □

This provides the following testable implication: if the percent change in toxic viewing is distinct from percent change in toxic sharing, users do not behave mechanically. In particular, the equality of these ratios implies that the elasticity of toxic sharing with respect to toxic viewing equals 1.

Table 3 presents the estimates from a serially unrelated regression (SUR) that stacks the two main outcomes: average toxicity in posts viewed and shared. This table also shows that the Trending Feed treatment significantly lowers the average toxicity in both views and shares. The stacked regression allows me to test non-linear hypothesis in Proposition 1.

I cannot reject that $\frac{\lambda_1^{bf}}{\lambda_0} = \frac{\pi_1^{bf}}{\pi_0}$ with a p-value of 0.283. However, for trending feed I observed that the percentage change in toxic views is $\frac{\lambda_1^{tf}}{\lambda_0} = 2.9\%$ and $\frac{\pi_1^{tf}}{\pi_0} = 3\%$. Therefore, I cannot reject the hypothesis that users behave mechanically when assigned to either

[9] This is referred to as the Overton Window in the Political Science literature, which represents the range of ideas that are acceptable in public discourse. The Overton Window is likely to change as norms of discussion change. See `https://www.nytimes.com/2019/02/26/us/politics/overton-window-democrats.html`.

treatment arms. That is, in both cases, the elasticity of toxic sharing with respect to toxic viewing cannot be distinguished from 1.

5 Model

The model tests for convex returns to sharing toxic content with respect to toxic exposure.

5.1 A Model of Sharing

I write down a simple model of user engagement with toxic content where engagement with such content depends on cumulative exposure of a user to toxic content in the past. Suppose user i is exposed to post p with a toxicity score of $t_{pi} \in [0, 1]$ on the platform. I postulate that the probability of user i sharing post p not only depends on the toxicity of post p, but also the average toxicity of all posts that the user has been exposed to in the session.

Then, suppose S_{pi} is an indicator function, taking value 1 when user i shares a toxic post p. The probability of sharing toxic posts can be written as,

$$\mathrm{E}[S_{pi}|t_{pi}, \bar{t_i}] = \beta_0 + \beta_1 t_{pi} + \beta_2 t_{pi} \cdot \bar{t_i} \tag{4}$$

where, $\bar{t_i}$ is the average of toxicity scores that user i was exposed to on their feed. The main parameter of interest is β_2 since I am interested in the probability of sharing a post, given it's toxicity score and a history of exposure to posts with varying degrees of toxicity. Assuming, $\beta_1 > 0$, I hypothesize that a history of higher exposure to toxic content is likely to increase the probability of sharing a toxic post, that is $\beta_2 > 0$.

This generates a testable implication that if average toxicity of user feed has a multiplier effect on the probability of sharing a toxic post, $\beta_2 > 0$. This measures the extent to which a user's past exposure to toxicity influences how likely they are to share a post, for given toxicity of a post p. Due to computational constraints, I estimate the linear probability model (4) at the user level. Therefore, averaging the toxicity of posts at the user level, I obtain

$$\mathrm{E}[\bar{S}_i|\bar{t}_i, \bar{t}_i^2] = \beta_0 + \beta_1 \bar{t}_i + \beta_2 \bar{t}_i^2 \tag{5}$$

where, $\bar{S}_i$ is the fraction of toxic posts shared by user i out of all the posts viewed by them. I test if $\beta_2 > 0$ which implies that social media mimics "winner-takes-all" markets. In other words, I check whether viewing more toxic content increases the toxicity in shares more than proportionately, as would be predicted by popular models of consumption on social media platforms (Salganik et al., 2006).

Estimated β_1 and β_2 are likely to be biased due to omitted variables, such as user preferences. That is, users who are shown posts with higher toxicity on average may also be users who are more likely to share (or in general, engage with) content because they spend a lot of time on the platform. In order to estimate the causal effect of toxicity of user feed on the fraction of posts shared, I use experimental variation in users' exposure to toxic feeds.

5.2 Instrumental Variables

I instrument toxicity of user feed using the local news experiment, as treatment assignment was shown to be random in Table 1. That is, I randomly vary $\bar{t}_i$ by displacing toxic posts with local news posts as treated users continue spending a fixed amount of time on the platform. I estimate the following first stage equation for the average toxicity of a user's feed,

$$\bar{t}_i = \gamma^0 + \gamma^{bf} D_i^{bf} + \gamma^{tf} D_i^{tf} + \varepsilon_i^{\pi} \tag{6}$$

where, D_i's are dummies for treatment assignment into the Bucket Feed or Trending Feed treatment arms. Assuming that users do not change the total number of posts they view on the platform, or their logging behavior, in response to treatment, I can test if treatment assignment in the local news experiment decreases toxicity of user feed on average, i.e. whether $\pi^{bf} < 0$, and $\pi^{tf} < 0$.

I am interested in estimating the rate at which exposure changes sharing behavior, while allowing for this rate to differ for different levels of average toxicity in user feeds. Therefore, using the two treatment arms as two instruments, I can also estimate the second raw moment of toxicity in user feeds,

$$\bar{t}_i^2 = \delta^0 + \delta^{bf} D_i^{bf} + \delta^{tf} D_i^{tf} + \varepsilon_i^{\delta} \tag{7}$$

where, D_i's denote random treatment assignment. Then, the structural equation estimating the effect of variance and average toxicity of user feeds on aggregate success of toxic posts on the platform is given by the following regression of fraction of posts shared by user i

$$\mathrm{E}[\bar{S}_i | \hat{t}_i, \hat{t}_i^2] = \beta_0^{IV} + \beta_1^{IV} \hat{t}_i + \beta_2^{IV} \hat{t}_i^2 + \mu_i \tag{8}$$

where, $\hat{t}_i$ and $\hat{t}_i^2$ are estimated from first stage regressions in (6) and (7). These IV estimates are identified under the exclusion restriction that treatment affects fraction of shares by user i only through the channel of average toxicity of user feed as well as it's second raw moment. That is to say, treatment assignment does not directly affect overall platform usage.

Then, I test the following hypothesis to check if the distribution of previous exposure to toxic content has a positive multiplier effect on sharing probabilities, that is $\beta_2^{IV} > 0$. The model enables me to test the hypothesis that toxicity of user feed has increasing returns in terms of fraction of posts that a user engages with, in the aggregate.

6 Results

I bring the testable implications from the model to the data.

6.1 OLS

I evaluate the structural relationship in (4) at the average toxicity of the feed to estimate the fraction of posts shared that are toxic. I reject the hypothesis that $\beta_2 > 0$ as Table C.2 shows that in a regression of toxic engagement (shares) on the first two moments of toxicity in user feed (views), the coefficient on toxicity squared is statistically significant and negative. This means that there are decreasing returns to sharing toxic posts viewed.

However, these estimates may be biased because average exposure to toxic content is endogenously determined by user's preferences for the platform and toxic content. These preferences may in turn predict the outcome of interest, i.e. the toxicity of shared content. This is because the first two moments of the toxicity of a user's feed are endogenously determined by user preferences as learnt by the personalization algorithm.

6.2 First Stage

I estimate equations (6) and (7) to test the strength of the first stage relationship between the two endogenous variables (average toxicity of feed and average toxicity of feed squared), and the two instruments (random treatment assignment to the bucket and trending feed arms). Table C.3 shows that treatment assignment strongly predict the first two moments of the distribution of average toxicity in views. In fact, this table shows that local news posts are likely to have displaced toxic posts from user feeds.

6.3 IV Estimates

I present IV estimates from (8) in Table 4. None of the coefficients are statistically significant, so I cannot reject the hypothesis that $\beta_1^{IV} = \beta_2^{IV} = 0$.

This implies that not only is the relationship between toxic views and shares linear, changing toxic views does not significantly alter sharing behavior either. This supports the result that users behave mechanically. Behavioral responses to this intervention are therefore, constant such that users share a fixed fraction of toxic posts they view irrespective of the change in exposure.

7 Conclusion

This paper provides evidence that a viewpoint-blind, algorithmic intervention–the insertion of localized, neutral news content–can reduce engagement with harmful discourse on social media platforms. The experiment conducted on a large-scale Indian platform demonstrates that exposing users to neutral local news reduces their interaction with divisive and toxic content, without requiring heavy-handed moderation practices like deplatforming or content removal. I show that the effect on sharing behavior is mechanical as the intervention also reduced exposure to toxic content in the same proportion.

The methodological contribution of this paper advances our understanding of user behavior on social media by devising a novel test to check if user sharing behavior is mechanical. I find that the benefits of this nudge-type intervention does not have a multiplier effect because users always share a fixed proportion of toxic posts they view. This is especially true when a user's content feed is not made to change drastically.

From a policy perspective, these results underscore the potential for viewpoint-blind interventions as a promising alternative to traditional content moderation approaches, which often provoke criticism for perceived censorship. By focusing on reducing the visibility of divisive content without explicitly targeting specific viewpoints, platforms can promote healthier online environments while avoiding accusations of bias or suppression. Even though the intervention does not have the kind of "snowballing" effects hoped for, a mechanical decrease in toxic sharing makes the intervention successful.

Local news reporting has been recognized for its potential to curb not just misinformation, but also to divert attention from mainstream news narratives which may be captured by governments and powerful interest groups (Petrova, 2008). This is especially important in India where the role of Indian TV news channels in furthering powerful political agendas has been documented (Garimella and Datta, 2024). However, there have been strong grassroots movements to offer local news as meaningful alternatives to these biased media sources. Two examples in the traditional media space are *Khabar Lahariya* and *Lankesh Patrika*.[10] Digital initiatives have also taken shape, for instance Way2News app raising $14 million in funding, with a business model that relies on local advertising Rag (2024).

Overall, this study contributes to the growing literature on content moderation, me-

[10] *Khabar Lahariya*'s contributions to society are well documented in the Oscar-winning film "Writing with Fire". More information can be found here: `https://khabarlahariya.org/about-us/`

dia bias, and social media's societal impact, particularly in rapidly growing yet scantily regulated markets like India. As social media platforms continue to expand and influence global discourse, understanding how algorithmic interventions can promote more civil and less harmful interactions is crucial. This is especially true in developing countries where regulatory frameworks, and their enforcement, may be weak and nascent even though the rate of adoption is very high.

This study suffers from limitations that carve a path for future research work. First, this type of intervention is expensive as it requires platforms to build an infrastructure to gather local news stories from local stringers. For this reason, the intervention was cut short in my setting as it was not sustainable due to low demand for local news and high costs of retaining journalists on the payroll. Second, my study only analyzes effects of the intervention in the short run. I do not expect the effects to persist as the short-run effects are small and the treatment was cut short due to high costs of the intervention.

Third, this nudge-intervention provided an extremely light-touch intervention. This has the advantage that changes to the feeds may not be fully perceived by users, thus making the mechanical sharing channel all the more plausible. However, while regulators search for view-point-blind policies, it is conceivable that real-world applications require stronger nudges. Four, this paper studies a particular platform in a specific context. The findings therefore, capture the effects of the intervention for given patterns in viewership and feed-ranking algorithms.

In this context, local news was shown to be politically neutral, which may not be the case in other places, such as the US. It is therefore, imperative that different view-point-blind settings are introduced across varied contexts, and tested for effectiveness accordingly. Further research should explore the long-term implications of such interventions and their applicability to other contexts and platforms.

References

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

H. Allcott, M. Gentzkow, and L. Song. Digital addiction. *American Economic Review*, 112 (7):2424–63, 2022.

C. Angelucci, J. Cagé, and M. Sinkinson. Media competition and news diets. *American Economic Journal: Microeconomics*, 16(2):62–102, 2024.

G. Aridor, D. Gonçalves, D. Kluver, R. Kong, and J. Konstan. The economics of recommender systems: Evidence from a field experiment on movielens. *arXiv preprint arXiv:2211.14219*, 2022.

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

E. Ash and S. Hansen. Text algorithms in economics. *Annual Review of Economics*, 15(1): 659–688, 2023.

S. Banaji, R. Bhat, A. Agarwal, N. Passanha, and M. Sadhana Pravin. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. 2019.

G. Beknazar-Yuzbashev, R. Jiménez Durán, J. McCrosky, and M. Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*, 2022.

L. Braghieri, R. Levy, and A. Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022.

K. Carney. The effect of social media on voters: experimental evidence from an indian election. *Job Market Paper*, 2022:1–44, 2022.

J. Chen and J. Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936, 2024.

C.-F. Chiang and B. Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of economic studies*, 78(3):795–820, 2011.

S. Dash, A. Arya, S. Kaur, and J. Pal. Narrative building in propaganda networks on indian twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 239–244, 2022.

S. DellaVigna and E. Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

M. Duggan. 5 facts about online harassment. 2014.

K. Garimella and A. Datta. Unraveling the dynamics of television debates and social media engagement: Insights from an indian news show. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 435–447, 2024.

M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

S. González-Bailón, D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. M. Guess, et al. Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398, 2023.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

H. Hosseinmardi, A. Ghasemian, A. Clauset, D. M. Rothschild, M. Mobius, and D. J. Watts. Evaluating the scale, growth, and origins of right-wing echo chambers on youtube. *arXiv preprint arXiv:2011.12843*, 2020.

D. Hume, J. Gathergood, and N. Stewart. The limits of nudge: Evidence from online property listings. *Available at SSRN 4846383*, 2024.

R. Jiménez Durán. The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN*, 2022.

A. Kalra. Hate in the time of algorithms: Evidence from a large-scale experiment on online behavior. Technical report, 2024. URL `https://github.com/aarushirita/aarushirita.github.io/raw/master/_pages/algorithms-jmp-KALRA.pdf`.

M. Katsaros, K. Yang, and L. Fratamico. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 477–487, 2022.

S. D. Kominers and J. M. Shapiro. Content moderation with opaque policies. Technical report, National Bureau of Economic Research, 2024.

R. Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870, 2021.

G. J. Martin and A. Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599, 2017.

K. Müller and C. Schwarz. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167, 2021.

K. Müller and C. Schwarz. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3):270–312, 2023.

M. Petrova. Political economy of media capture. *Information and public choice*, page 121, 2008.

N. Purnell and R. Roy. Facebook faces hate-speech questioning by indian lawmakers after journal article. *Wall Street Journal*, 2020. URL `https://www.wsj.com/articles/facebook-faces-hate-speech-grilling-by-indian-lawmakers-after-journal-article-11597747734`.

A. Rag. Hyperlocal news startup way2news raises $14 million in funding from existing backer westbridge, others. 2024. URL `https://economictimes.indiatimes.com/tech/funding/hyperlocal-news-startup-way2news-raises-14-million-in-funding-from-westbridge-capital-others/articleshow/113420783.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst`.

S. G. Roy, U. Narayan, T. Raha, Z. Abid, and V. Varma. Leveraging multilingual transformers for hate speech detection. *arXiv preprint arXiv:2101.03207*, 2021.

M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

J. M. Snyder Jr and D. Strömberg. Press coverage and political accountability. *Journal of Political Economy*, 118(2):355–408, 2010.

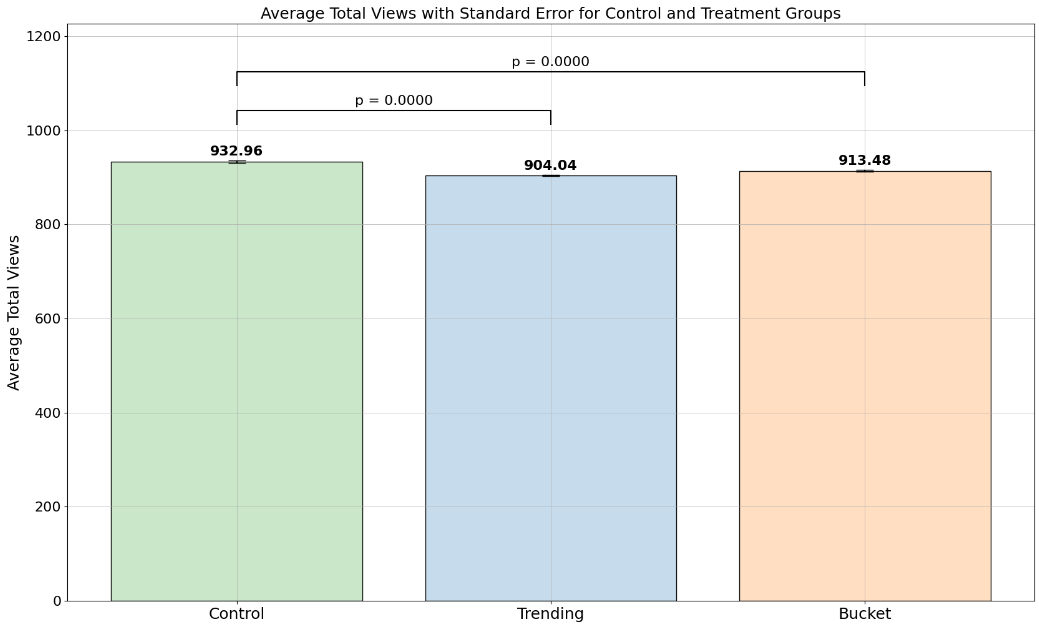
Statista. Daily time spent on social networking by internet users worldwide from 2012 to 2023 (in minutes). Technical report, 2023. URL `https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/`.

N. Thakral and L. T. Tô. *When Are Estimates Independent of Measurement Units?* Boston University-Department of Economics, 2023.

R. H. Thaler and C. R. Sunstein. *Nudge: The final edition.* Yale University Press, 2021.

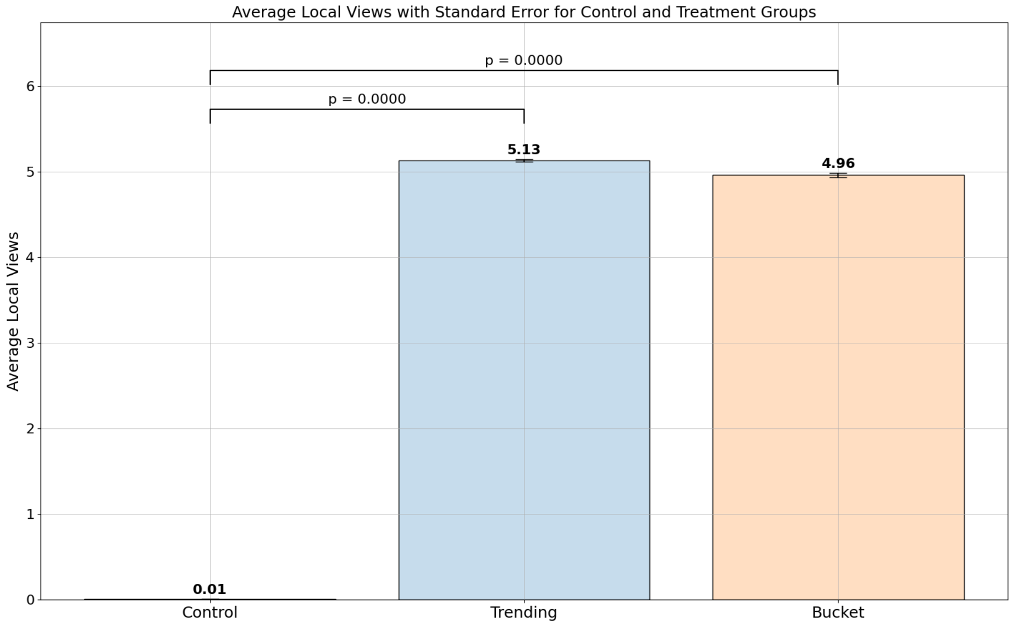
Tables and Figures

Figure 1: Experimental effects on total number of posts viewed



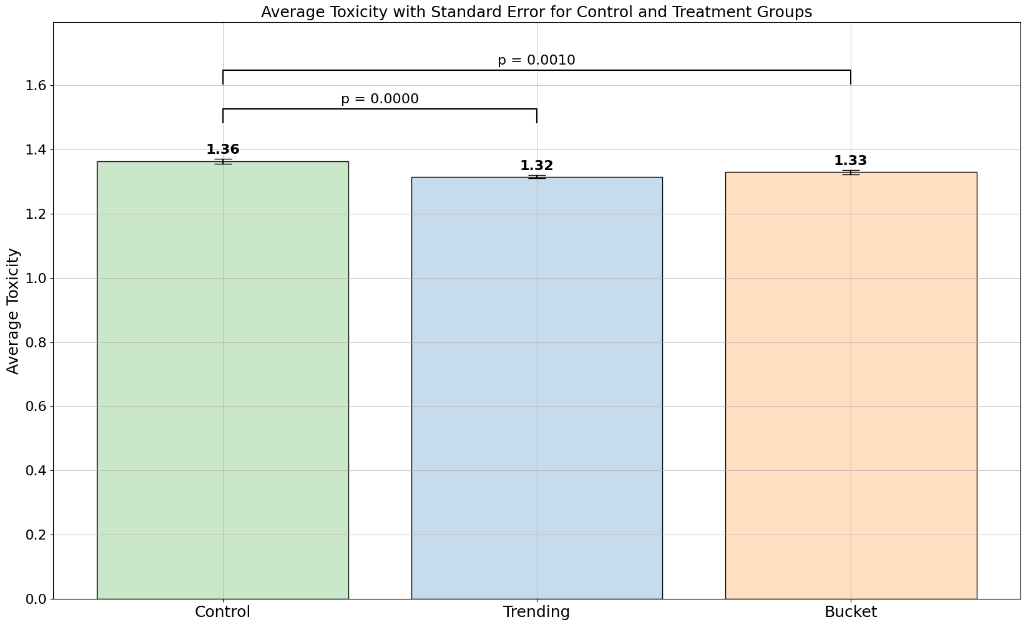
Notes: This figure shows that there are no significant differences in the total number of posts viewed by users assigned to different treatment arms, during the intervention period of March 2021. Although these differences in average number of posts viewed is statistically significant, I show that these magnitudes are economically insignificant.

Figure 2: Effects on local news posts viewed by treatment assignment



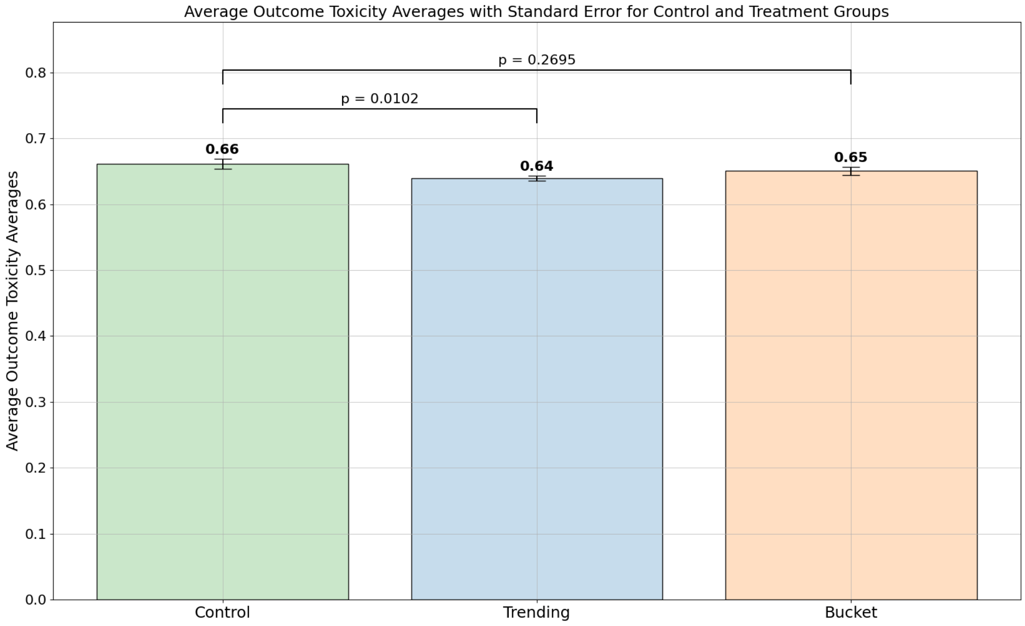
Notes: This figure shows that although the intervention induced a strong "first stage" effect on users' exposure to local news posts, the intervention provides a small nudge in content consumption away from toxic posts. This means that the main results are ITT estimates. The exposure to local news is interpreted as treatment intensity as it displaces exposure to toxic posts.

Figure 3: Experimental effects on average toxicity of content feed



Notes: This figure shows that the intervention reduces exposure to toxic content. The decrease in the average toxicity of content feeds is larger for users assigned to the trending feed treatment arm. Data is aggregated at the user level and robust standard errors are computed.

Figure 4: Experimental effects on average toxicity of shares



Notes:

Table 1: Verifying validity of randomization in treatment assignment across observable user characteristics

Characteristic	BucketFeed	TrendingFeed
city: dindigul	-0.035	0.039
	(0.025)	(0.03)
city: fatehgarh_sahib	0.011	-0.002
	(0.021)	(0.025)
city: ramgarh	0.008	0.03
	(0.019)	(0.021)
city: chikballapur	-0.001	0.037
	(0.03)	(0.035)
city: malda	0.052	-0.018
	(0.04)	(0.044)
city: ranchi	0.012	0.007
	(0.014)	(0.016)
city: sonepat	-0.002	0.026
	(0.014)	(0.016)
city: jamtara	-0.009	0.051
	(0.029)	(0.034)
city: mahisagar	0.002	0.04
	(0.034)	(0.039)
city: uttara_kannada	-0.015	0.002
	(0.02)	(0.024)
city: east_singhbhum	0.01	0.011
	(0.015)	(0.017)
city: beed	0.003	0.011
	(0.014)	(0.016)
ageRange: above 40	-0.236	0.39
	(0.016)	(0.016)
city: madurai	-0.013	0.006
	(0.022)	(0.026)
city: jhalawar	0.01	-0
	(0.015)	(0.017)
city: valsad	-0.007	0.021
	(0.014)	(0.017)
gender: Male	-0.001	0.002
	(0.001)	(0.001)
ageRange: 31-45	0.006	-0.012
	(0.013)	(0.015)
Number of observations:	1312535	1312535

Notes: This Figure shows that treatment assignment is balanced in (a randomly selected set of) observable user characteristics. I cannot reject the null hypothesis that these characteristics jointly determine treatment assignment with a p-value greater than 0.1.

Table 2: Summary statistics for experimental sample

Metric	Control	BucketFeed	TrendingFeed
Total Views (December 2020)	1129.789	1120.361	1110.941
Local News Views (December 2020)	43.374	42.52	42.183
Average posts supplied in March 2021	10266.702	10264.147	10239.286
User Number	210,824	313,944	784,127

Notes: This table shows that the the platform produced close to 10,000 local news posts on average in the locations clusters that users in the experimental sample belong to. There are a total of 1.4 million users in the balanced sample across the treatment arms.

Table 3: Seemingly Unrelated Regressions (SUR) of average toxicity in views and shares

Average Toxicity in Views (%)	
Bucket Feed	-0.026**
	(0.010)
Trending Feed	-0.043***
	(0.009)
Control Mean	1.363***
	(0.008)
Average Toxicity in Shares (%)	
Bucket Feed	-0.009
	(0.010)
Trending Feed	-0.021*
	(0.009)
Control Mean	0.661***
	(0.008)
N	1312535

Notes: This table shows the results of an SUR stacking two outcomes: the average toxicity of views and shares. It finds that there is a significant reduction in the toxicity of views and shares for users randomly assigned to the Trending Feed treatment arm. The stacked regression design enables testing of nI cannot reject that $\lambda_1^{bf}/\lambda_0 = \pi_1^{bf}/\pi_0$ with a p-value of 0.283. However, for trending feed I observed that the percentage change in toxic views is $\lambda_1^{tf}/\lambda_0 = 2.9\%$ and $\pi_1^{tf}/\pi_0 = 3\%$. Therefore, I cannot reject the hypothesis that users behave mechanically when assigned to either treatment arms. These ae also the main reduced form estimates. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: IV estimates tracing the shares and toxic shares as function of toxicity of user feed

	(1) Average toxicity in shares
Average toxicity in views	1.236 (1.727)
Average toxicity in views squared	-0.046 (0.109)
Constant	-0.327 (0.752)
N	1312535

Notes: Outcome variable is aggregated at user-month level for the intervention period in March, 2021. Robust standard errors in parenthesis. $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A Proofs

A.1 Proof of Proposition 1

Consider a special case of the model so that sharing behavior does not respond to treatment assignment or the exposure to toxic content during the intervention period. Further the probability of sharing toxic content, for given levels of toxic exposure, is constant across all users, so that all users behave mechanically. Then the proportionate change in sharing toxic content is exactly equal to the proportionate change in toxic content viewed, i.e.

$$\frac{\lambda_1^{bf}}{\lambda_0} = \frac{\pi_1^{bf}}{\pi_0} \text{ and } \frac{\lambda_1^{tf}}{\lambda_0} = \frac{\pi_1^{tf}}{\pi_0}$$

Proof. Without loss of generality, consider one treatment arm, bf or tf, and denote treatment status using D_i. Suppose each individual i sees a fixed number of posts J.

Let $s_{ij} \in \{0,1\}$ indicate the individual's sharing of post $j = 1, \ldots, J$. Let $t_{ij} \in \{0,1\}$ indicate the toxicity of post $j = 1, \ldots, J$ seen by individual i. Then, define the toxicity score as

$$T_i = \frac{1}{J} \sum_j t_{ij}$$

the sharing rate as

$$S_i = \frac{1}{J} \sum_j s_{ij}$$

and the toxicity of shares as

$$Y_i = \left(\sum_j s_{ij} t_{ij} \right) / \left(\sum_j s_{ij} \right)$$

Suppose there is no treatment effect on sharing, just on toxicity. So we have potential outcomes $t_{ij}(1)$ and $t_{ij}(0)$ but s_{ij} is unaffected by treatment D_i. Then regressing Y_i on randomly assigned D_i identifies a slope coefficient of

$$E\left[Y_i(1) - Y_i(0)\right] = E\left[\frac{\sum_j s_{ij}(t_{ij}(1) - t_{ij}(0))}{\sum_j s_{ij}} \right] \equiv \pi_1 \tag{9}$$

and an intercept of

$$E[Y_i(0)] = E\left[\frac{\sum_j s_{ij} t_{ij}(0)}{\sum_j s_{ij}} \right] \equiv \pi_0 \tag{10}$$

Consider also the regression of T_i on D_i: it yields a slope coefficient of

$$E[T_i(1) - T_i(0)] = E\left[\frac{1}{J} \sum_j (t_{ij}(1) - t_{ij}(0)) \right] \equiv \lambda_1 \tag{11}$$

and an intercept of

$$E[T_i(0)] = E\left[\frac{1}{J}\sum_j t_{ij}(0)\right] \equiv \lambda_0 \tag{12}$$

The claim is now that $\pi_1/\pi_0 = \lambda_1/\lambda_0$. This requires the following assumptions:

1. *Monotonicity*: $t_{ij}(1) \leq t_{ij}(0)$ for all i, j
2. *Nonresponsiveness*: $Pr(t_{ij}(1) = 0 \mid t_{ij}(0) = 1, s_{ij}) = \delta$ for all i, j

The first assumption says toxicity can only decrease in the experiment; the second assumption says the chance of this decrease is unrelated to the baseline share status s_{ij} and is moreover constant over individuals and posts. Under these two assumptions

$$t_{ij}(1) = (1 - p_{ij})t_{ij}(0)$$

where, $E[p_{ij} \mid t_{ij}(0), s_{ij}) = \delta$, so $t_{ij}(1) - t_{ij}(0) = -p_{ij}t_{ij}(0)$ and

$$\begin{aligned}
\pi_1 &= E\left[\frac{\sum_j s_{ij}(t_{ij}(1) - t_{ij}(0))}{\sum_j s_{ij}}\right] \\
&= E\left[\frac{-\sum_j s_{ij}p_{ij}t_{ij}(0)}{\sum_j s_{ij}}\right] \\
&= E\left[E\left[\frac{-\sum_j s_{ij}p_{ij}t_{ij}(0)}{\sum_j s_{ij}} \mid s, t(0)\right]\right] \\
&= -\delta \cdot E\left[\frac{\sum_j s_{ij}t_{ij}(0)}{\sum_j s_{ij}}\right] \\
&= -\delta\pi_0
\end{aligned}$$

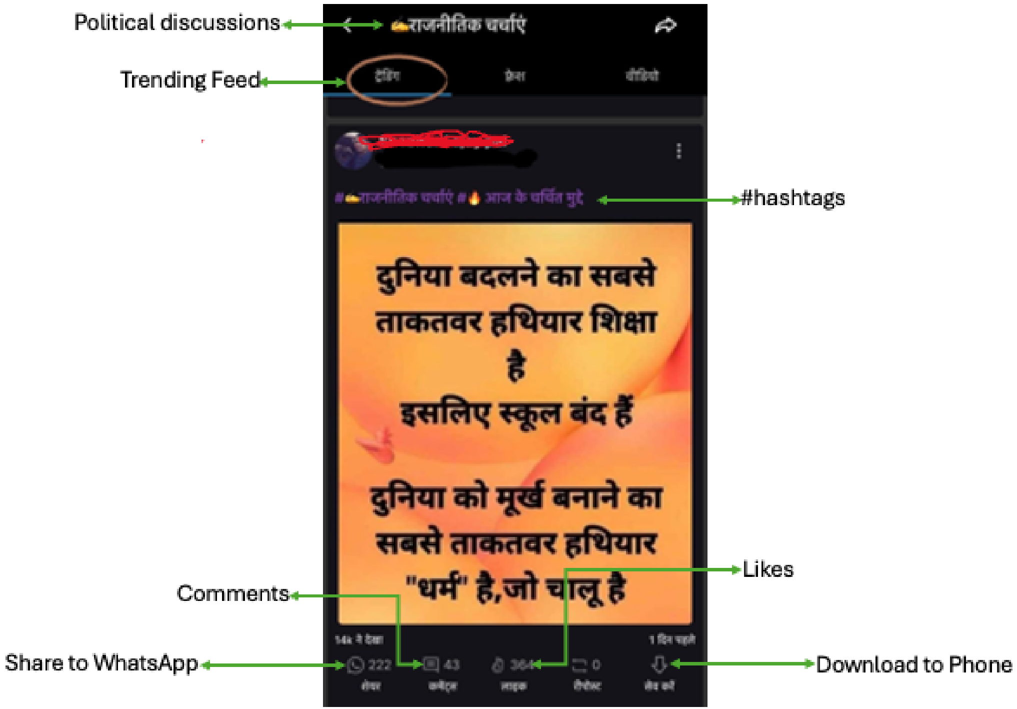
using the model and the LIE. Moreover,

$$\begin{aligned}
\lambda_1 &= E\left[\frac{1}{J}\sum_j (t_{ij}(1) - t_{ij}(0))\right] \\
&= -E\left[E\left[\frac{1}{J}\sum_j p_{ij}t_{ij}(0) \mid s, t(0)\right]\right] \\
&= -\delta E\left[\frac{1}{J}\sum_j t_{ij}(0)\right] \\
&= -\delta\lambda_0
\end{aligned}$$

Thus, $\pi_1/\pi_0 = \lambda_1/\lambda_0 = -\delta$. □

B Supplementary Figures

Figure B.1: SM landing page and trending feed

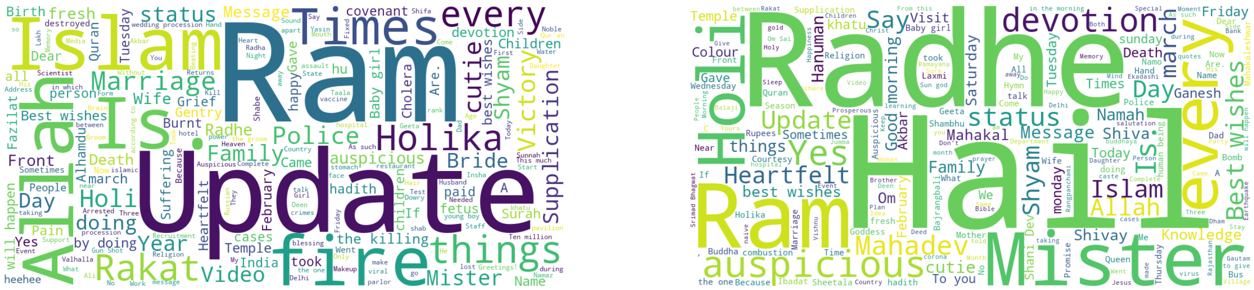


दुनिया बदलने का सबसे
ताकतवर हथियार शिक्षा
है
इसलिए स्कूल बंद हैं

दुनिया को मूर्ख बनाने का
सबसे ताकतवर हथियार
"धर्म" है,जो चालू है

Notes: This Figure shows the landing screen or trending feed on SM, and an example of an image post on the platform. Users in the bucket feed treatment arm saw local news posts in a separate tab, that may be accessed by swiping towards the left on this home screen. This also shows the different ways users can choose to engage with the post. Of primary importance is the feature which enables users to directly share the post with their friends on WhatsApp.

Figure B.2: Word clouds depicting words associated with highly toxic posts

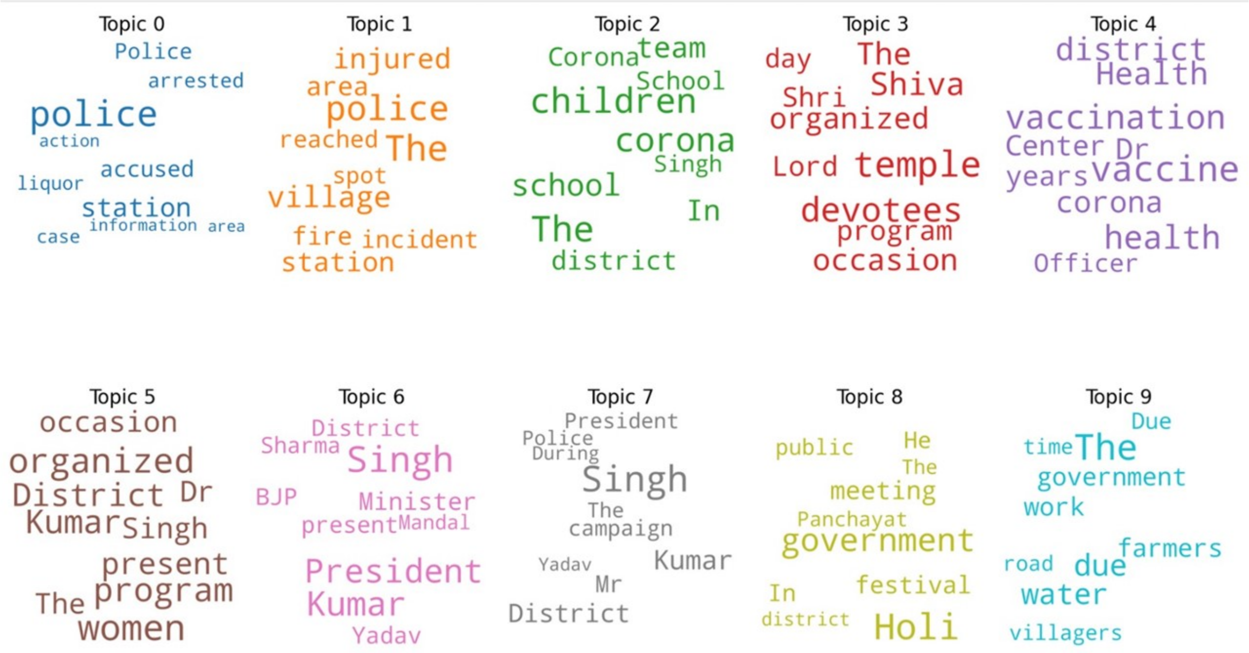


(a) High Toxicity

(b) Low Toxicity

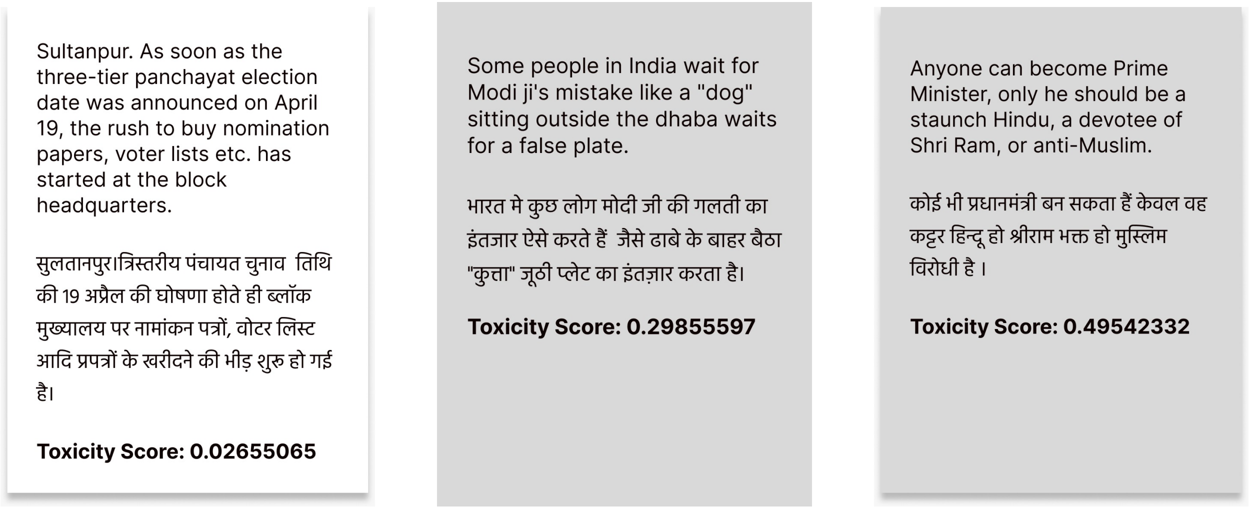
Notes: This Figure shows word clouds constructed using the TF-IDF vectorizer, on posts classified into high and low toxicity categories respectively. Cut-off to classify posts into high and low toxicity categories is 0.2, based on the toxicity scores provided by Perspective API. The figure demonstrates overlap in words pertaining to religion in both categories, for example 'Islam' and the Hindu mythological god-king 'Ram,' who is also central to Hindu nation building agenda of the current ruling government. This highlights the need for contextual embeddings to characterize the text data. Perspective's toxicity algorithm uses human labeled comments and BERT models to provide toxicity scores to each post, by representing posts in some latent space as embedding vectors. These posts were extracted from the platform in April 2023, and the cut-off in toxicity scores was employed for illustration purposes only.

Figure B.3: Topics in local news content



Notes: This Figure shows topics that most commonly appear in local news stories, using an LDA model. This shows that local news posts are politically neautral, providing readers facts about local crime, public meetings, and festivals. During the intervention period, users largely followed local news for updates on CoronaVirus cases in neighborhoods, and to find information on immunization drives.

Figure B.4: Examples of local news posts and political posts in Hindi



Sultanpur. As soon as the three-tier panchayat election date was announced on April 19, the rush to buy nomination papers, voter lists etc. has started at the block headquarters.

सुलतानपुर।त्रिस्तरीय पंचायत चुनाव तिथि की 19 अप्रैल की घोषणा होते ही ब्लॉक मुख्यालय पर नामांकन पत्रों, वोटर लिस्ट आदि प्रपत्रों के खरीदने की भीड़ शुरु हो गई है।

Toxicity Score: 0.02655065

(a) Local news post

Some people in India wait for Modi ji's mistake like a "dog" sitting outside the dhaba waits for a false plate.

भारत मे कुछ लोग मोदी जी की गलती का इंतजार ऐसे करते हैं जैसे ढाबे के बाहर बैठा "कुत्ता" जूठी प्लेट का इंतज़ार करता है।

Toxicity Score: 0.29855597

(b) Political post

Anyone can become Prime Minister, only he should be a staunch Hindu, a devotee of Shri Ram, or anti-Muslim.

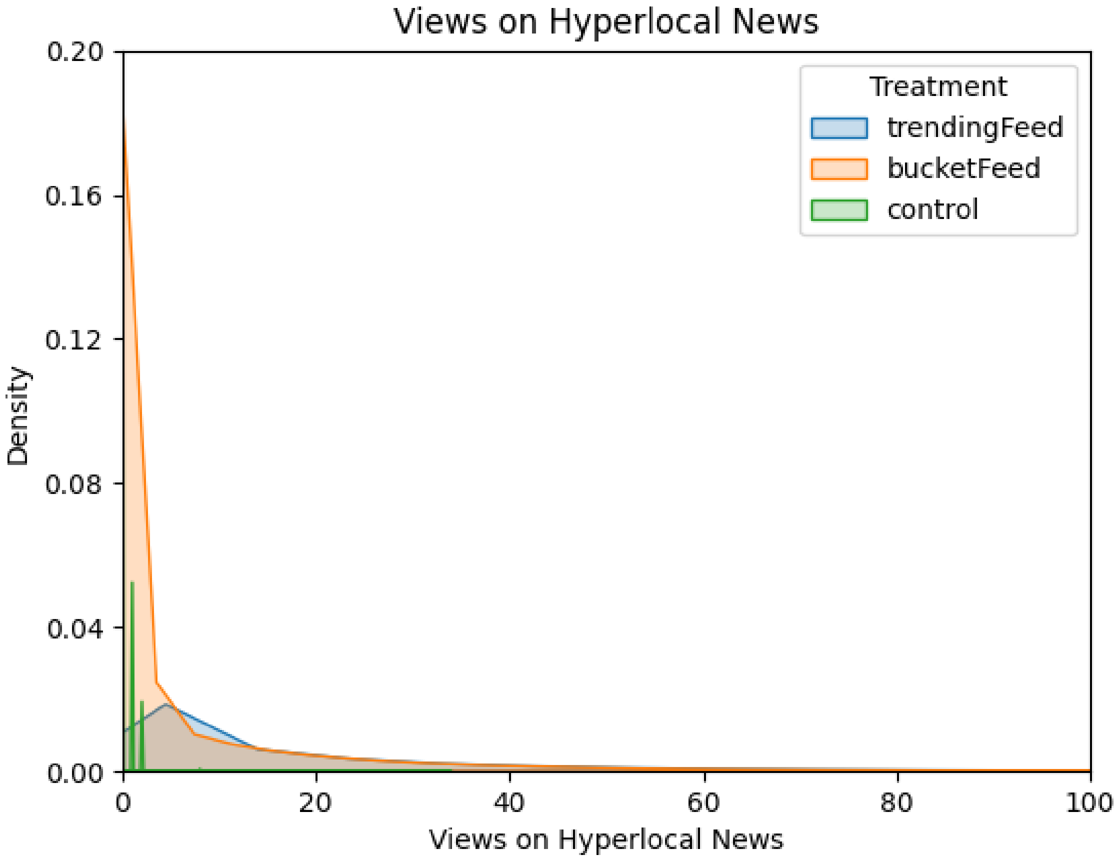
कोई भी प्रधानमंत्री बन सकता हैं केवल वह कट्टर हिन्दू हो श्रीराम भक्त हो मुस्लिम विरोधी है ।

Toxicity Score: 0.49542332

(c) Political posts with high toxicity

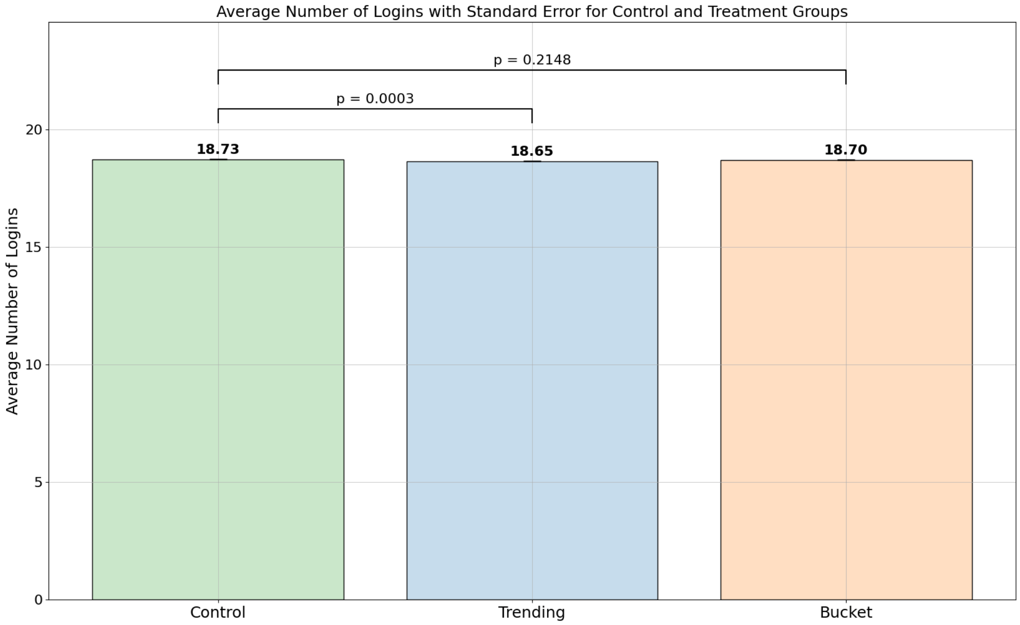
This Figure provides examples of toxicity scores generated by the Perspective algorithm. I provide Hindi text along with the English translation. To fix ideas, Panel (a) shows the toxicity score on a local news post in the data. I provide more examples with different toxicity scores in the Supplementary material.

Figure B.5: Distribution of exposure to local views across treatment arms.



Notes:

Figure B.6: Experimental effects on total number of logins



Notes:

C Supplementary Tables

Table C.1: First-stage results on number of local news posts viewed

Treatment	Local News Posts Views	Probability of Viewing Any Local News
Bucket Feed	4.952***	0.203***
	(0.023)	(0.001)
Trending Feed	5.123***	0.203***
	(0.015)	(0.000)

Notes: This table shows the first stage regression estimates provide effect of treatment status on exposure to local news posts. There is a statistically significant increase in exposure to local news content for all treatment arms, compared to the control group. Views on local news in column (1) are given in units of posts, and are expressed in probability terms in column (2). Outcome variables are aggregated at user-month level for March, 2021. Robust standard errors in parentheses.

Table C.2: OLS estimates of structural relationship between toxic views and shares

	Average toxicity in shares (%)
Average toxicity in views (%)	0.059***
	(0.002)
Average toxicity in views squared	-0.002***
	(0.000)
Constant	0.594***
	(0.003)
N	1312535

Notes: Outcome variable is aggregated at user-month level for the intervention period in March, 2021. Total shares are given by the total number of posts shared by a user in the month. Total toxicity on shared posts is the sum of toxicity scores (between 0 and 1) on all the posts shared by user. OLS regression estimates tracing the shares and toxic shares as function of toxicity of user feed. There is a statistically significant increase. Robust standard errors in parenthesis. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C.3: Strength of first stage in structural model of sharing

	(1) Average toxicity in views	(2) Average toxicity in views squared
Bucket Feed	-0.026**	-0.506*
	(0.010)	(0.210)
Trending Feed	-0.043***	-0.690***
	(0.009)	(0.185)
Constant	1.363***	15.142***
	(0.008)	(0.168)
N	1312535	1312535

Notes: This table shows the strength of the first stage relationship between the two endogenous variables and two instruments. I find that the instruments strongly predict both the first and second moments in the distribution of average toxicity of user feeds. Robust standard errors in parenthesis. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$