Algorithmic Drivers of Hate Speech on Social Media

Aarushi Kalra[*]

February 2023

Abstract

Social media algorithms are an increasing part of our everyday lives, yet little is know about the causal effect of these algorithms on individual well-being or on social welfare. In this project, first, I study the contribution of algorithmic recommendation systems on increased political polarization and anti-minority hate speech in India. Second, I study the effect of algorithms on consumer surplus and social welfare. This is done by disentangling the effects of user preferences for hate speech and biases against out-group members, from the effects of algorithmic amplification of hateful content in a large-scale experiment, in cooperation with one of India's largest social media platforms. I study the causal effect of these algorithms on user engagement with polarizing content on the platform, as well as on survey outcomes including users' subjective well-being, out-group bias and willingness to pay for content customization via algorithms.

[*] Department of Economics, Brown University, Providence, RI, 02906 (email: aarushi_kalra@brown.edu)

1 Introduction

Around the world, an increasing number of people are known to be spending greater periods of time on social media. Yet, the fact that users are increasingly reporting dissatisfaction with social media usage (Kleinberg et al., 2022) also points to evidence demonstrating that the increased use of social media may not be a result of consumer surplus, but rather is on account of habit formation and self-control problems (Allcott et al., 2022) created by the way social media platforms recommend content to its users, i.e. via algorithms. Algorithmic content recommender systems also generate externalities for other users in the networks on social media platforms, by altering the set of posts that they are exposed to, thus affecting social surplus adversely. Furthermore, these algorithms are also blamed for increased polarization and hate speech which affects both consumer and social welfare. These multiple downstream effects of algorithms on social and consumer welfare are, therefore, important to study in a context like India, which ranks only second to the US in social media consumption and yet remains a remarkably understudied context.

In this project, first, I study the contribution of algorithmic recommendation systems on increased political polarization and anti-minority hate speech in India. Second, I study the effect of algorithms on consumer surplus and social welfare. Content recommendation systems are algorithms that are customized according to user preferences to enhance content engagement on the platform. This is then, hypothesized to create rabbit holes where user discovery and user engagement with undesirable forms of content is very likely. I adopt the definition of 'rabbit holes' from (Piccardi et al., 2022), where internet rabbit holes are navigation paths followed by social media users that lead to long explorations, often about the original topic, and sometimes involving unexpected posts. I will test that rabbit holes create echo chambers where users with similar content preferences and characteristics continuously interact with each other, reinforcing negative beliefs about out-group members, or even changing such beliefs for the worse.

My main outcome variable is engagement with hate speech on SM, a content generation platform with millions of active monthly users in India. Like TikTok, SM is an upstream content generation app that relies on algorithmic content recommendation systems, and not on the network of friends or followers, to recommend images and videos to users. I cooperate with SM to conduct an experiment, randomizing over recommendation algorithms, that alters the set of posts appearing in users' feeds (See Section 4 for Experiment Design). Previous work on algorithmic content recommendation have provided useful insights about pathways to political polarization on social media platforms (Barberá et al., 2015; Conover et al., 2011), but consensus on the causal effect of algorithms on engagement with hateful and polarizing content is yet to be established. The absence of agreement on the causal relationship between algorithms and hate speech, in part, reflects limitations in available data, which is not well suited to measure consumption of different types of content over extended periods of time (Hosseinmardi et al., 2020), as well as the absence of experimental variation in algorithmic exposure.

1.1 Research Questions

While regulating platform algorithms may decrease polarization and hate speech, a platform designer may be faced with a trade-off, because such regulation can also decrease both platform profits and consumer surplus (if users have a taste for toxic, hateful, or polarizing content). The research questions are: *How do algorithms that recommend content to users on social media platforms contribute to:*

- Users' subjective well-being
- Users' engagement with polarizing and hateful content
- Users' engagement with their preferred types of content
- Digital addiction
- Users' willingness to pay for content personalization
- Attitudes towards out-group members

I study the downstream effects of social media rabbit holes induced by content recommender systems, or algorithms. I also test if effects on user behaviour are heterogeneous across users, and if this heterogeneity can be predicted ex-ante from user characteristics that are known to the platform.

2 Experiment Design

2.1 Intervention and Randomization

I cooperate with SM to conduct an experiment to measure the effects of algorithms on political polarization and hate speech. The control group consists of a random sample of users who are exposed to a ranked list of posts, where the ranking is determined by the user preferences as they are revealed in their previous engagement and are learnt by the algorithm. I describe the two-step process that produces content recommendations for SM users:

Candidate Generator: The recommendation system first creates a list of content pieces that are suitable candidates to be surfaced on the content feed. Typically, the CG creates a pool of 10,000 posts, from a corpus of 2 million posts in each language available on the platform, for each user every day based on relevance scores given to about 2 million posts in this process. The posts are personalized using the baseline characteristics of users that are known to the platform. These include user's gender and age, as well as post characteristics like tag genre of the post.

Ranker: The ranker then picks up the top 100 posts according to the relevance scores generated by the CG process and generates new scores according to previous engagement by the user with different kinds of content (where content type is understood using hash tags on posts). These new scores determine the rank of a post in the user feed.

Treated users are shown a list of content which is not ranked according to user preferences but are instead exposed to posts that are randomly drawn from a set of 'candidate' posts. I generate the treatment arms to construct appropriate counterfactuals to algorithmic customization, with varying degrees of content customization by altering the following dimensions of the recommendation system: **1)** Candidate Generator: I pick top 100, or all the 2 million posts from the CG each day for users assigned to different treatment arms; **2)** Ranker: For the given number of picks from the CG process, I then pick posts uniformly at random to populate the entire user feed.

2.2 Stratification

The intervention measures the treatment effect of varying degrees of customization in content exposure on user behavior while consuming content on social media platforms. However, this intervention will have spillover effects on other users that are connected to the treated user in their network on the platform. As a result, the estimated treatment effect is a function of the treatment status of other users. Therefore, in the absence of stratification, the estimated treatment effect is difficult to interpret because the control users in the network of the treated users would also be affected by the intervention.

To address this problem, and if it is possible, I will assign the treatment at the neighborhood level and not the user level because these neighborhoods form a good proxy for local content networks. Moreover, such stratification greatly bolsters the statistical power to analyze outcome variables determined at constituency or neighborhood level, like political participation in elections or protests, as well as results of elections. I will assign treatment to all users in 10,000 neighborhoods (in each treatment arm). With 200 users residing in an average Indian neighborhood, my sample consists of 2,000,000 users in each treatment arm, and 4,000,000 users in the control group. The combination of randomization across the two stages of the customization algorithm creates two treatment arms, as depicted in Table 1.

	Percent ranked feed randomized in treatment = 100%	Treatment	Control
Candidate Generator	100	2,000,000	4,000,000
	2000000	2,000,000	

Table 1: Number of users randomly picked in each variation of the experiment, and number of posts picked at candidate generator stage in the intervention.

The main analysis is conducted using the second row of Table 1, which has a higher 'degree of randomness,' or a lower degree of customization and content personalization. This is because all survey outcomes may not be available for users in the first treatment bin, as this intervention has already been run on SM as a 'pilot' intervention. Comparisons will be made with all users in the control group.

2.3 Data Collection

SM will begin piloting the intervention as designed above in late February 2023. User responses on the baseline survey will be collected in February, right before the intervention is implemented in March, 2023. Survey data collection will begin in March 2023, close to 4 weeks after the treated users were exposed to 'randomized' feeds. I expect to complete surveys with roughly 32,000 users (across treatment and control groups). These are platform based surveys, and the respondents are invited to participate in the survey via a text message sent to the users on the app itself. All users who are part of the sample survey will also receive notifications on their phones and by text message to fill out the questionnaire on the platform. This addresses concerns of *differential attrition* between treated and control users, especially to elicit responses for the endline survey.

Those respondents who opt in by responding to the message or clicking a link in the message will be sent to the landing page of the survey. Upon landing to this page, respondents will first be asked to read and sign a consent form that is linked in Appendix A in this document. The survey instrument (See Appendix B) consists of questions on demographics, subjective well-being, out-group attitudes, willingness to pay for the content customization, and satisfaction with the platform. I expect each survey to take roughly 5 to 10 minutes. Respondents will be compensated by transferring an incentive of INR 100 (in Amazon vouchers) upon completion of the survey. Data collection will continue until April 2023.

2.4 Definitions and Measurement

In the next section, I discuss primary hypothesis and measurement using survey outcomes. Here, I reiterate the definitions and measurement of some key terms and variables constructed from the administrative data.

- **Content Recommendation System/ Algorithm:** Content recommendation systems are algorithms that are customized according to user preferences to enhance content engagement on the platform. SM's Data Science and ML teams continually strive to optimize these algorithms to maximize user retention on the platform. SM uses data on previous engagement of a user with content, as well as static characteristics of users to generate user embeddings which helps predict the kind of content a user would like. This unsupervised machine learning algorithm is very similar to Netflix's automated movie recommendation system (Koren et al., 2009).

- **Hate Speech:** Anti-minority hate speech is the main outcome variable for the research questions I seek to answer. Since multilingual hate speech classification is a key part of my research, I develop an effective and automated pipeline to first translate millions of SM posts to English, and then using Perspective API to identify toxicity in the translated text. Perspective API, from Jigsaw and Google, provides the current best machine learning solution for toxicity detection, as it relies on training data "containing over one million toxic and non-toxic comments from Wikipedia," marked by human raters (Jiménez Durán, 2022). I label posts with a toxicity score higher than 0.3 as hateful. In Figure 1, I validate the performance of my method for multi-lingual hate speech detection by comparing results with the choice of hate speech classification algorithm and with manually annotated SM posts.

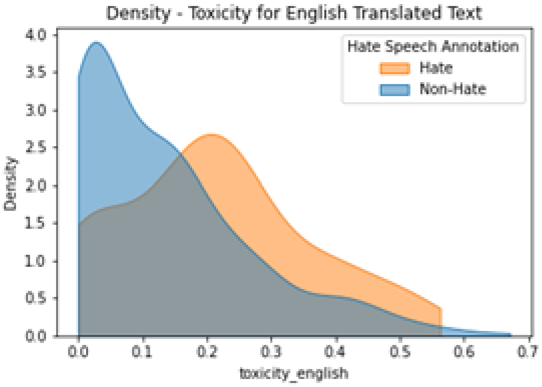


Figure 1: Density of toxicity scores from Perspective API by manual annotation of posts as hateful or not. Human annotators followed definition of hate speech as in De, et al. (2021).

- **Rabbit Holes:** I adopt the definition of 'rabbit holes' from Piccardi et al. (2022), where internet rabbit holes are navigation paths followed by social media users that lead to long explorations, often about the original topic, and sometimes involving unexpected posts. I call a user login session a rabbit hole if more than 30 to 40% of posts viewed in that user-session come from the same 'topic' (where topics in the text data are modelled following Gentzkow et al. (2019)).

- **Content Network:** The data environment provides a rich network of users and content, represented with a graph G. This graph consists of users, that belong to the set of nodes, $i \in \mathcal{I}$. These nodes (or users) are connected to each other via edges, $e \in \mathcal{E}$. An edge between two users i and j exists if they interacted with a common piece of content. The edges are weighted by the number of common pieces of content. $w_{i,j} = \#$ of common pieces of content users i and j interacted with, is the weight of the edge between users i and j.

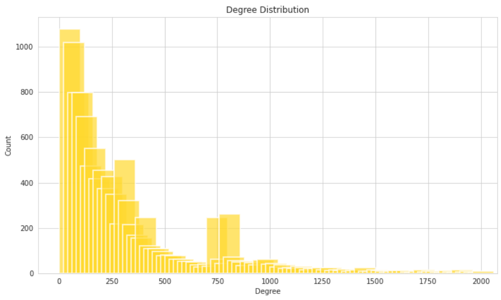


Figure 2: Degree Distribution for graph constructed using engagement with political content Hindi-speaking users in Uttar Pradesh, India.

> As an example, I summarize the structure of the network of Hindi users in Uttar Pradesh, centred around the topic of Politics. I have a disconnected graph with 5 components, with average degree around 360. Figure 2 presents the degree distribution of this graph. User engagement on the platform tends to be localized with respect to users physical neighbourhood. I observe a high degree of clustering in the network, where the magnitude of global clustering coefficient (Opsahl, 2013) is 35%. Modal local clustering coefficient (Watts and Strogatz, 1998) in this network is 1. This indicates that every neighbour connected to node $i \in \mathcal{I}$ is also connected to every other node in that neighbourhood .

Remaining measurement issues are addressed using the exact wording of the survey questionnaire in the next section.

3 Primary Hypothesis and Outcomes

With the experimental variation described above, I seek to answer the following research questions: ***1)*** *Do social media platforms create Internet Rabbit Holes resulting in increased Hate Speech and Political Polarization?* ***2)*** *What are the welfare implications of social media use in the world's second largest market for such digital products, in the presence of algorithmically induced, hateful rabbit holes?* I examine these questions by considering on-platform outcomes (like post intervention engagement with hate speech) and off-platform outcomes that are measure with user surveys. The survey data complements observational data collected by the platform as it provides rich information about user types, and helps the researcher identify how algorithms drive online behaviors differently by social group and other demographic characteristics of the users.

For each family of outcomes, I highlight the specific hypotheses (in bold text) along with the relevant questions from the survey instrument below (precise wording of each question in italics). In several cases, I have multiple outcomes of interest for each hypothesis. I pool these outcomes into a single test by constructing an "outcome index" that is the average of z-scores of all the outcome variables associated with that hypothesis. Thus, the primary variables of interest for each hypothesis will be the respective outcome indices. I discuss the empirical strategy in Section 4, along with corrections for multiple hypothesis testing.

3.1 Attitudes

In the **attitudes** family of outcomes, I test two hypothesis:

1. **Change in perception of out-group members.** I gauge the respondent's preferences over members of *other* groups by asking the following questions.

    - *Now, I would like to learn your preferences over content. For the following pairs of posts, please select only one option.*

      Which of the following posts would you like or share or comment on, if it came up on your feed.

    - ¡Image Description¿ *Prime Minister Modi's approval rating surpasses 22 global leaders, including Biden and Sunak.*
    - ¡Image Description¿ *Price of Milk has been rising.*

2. **Change in preference over content by neighbourhood or location.** Residential neighborhoods in India are highly likely to be segregated across religious and caste lines (Kalra, 2021). The following question elicits user preferences over content from own neighborhood versus content that is more global in nature.

    - *Choose the physical radius you would like to see news from:* (select multiple)
        - *Your neighbourhood or local settlement*
        - *Your city*
        - *Your state*
        - *All India*
        - *USA*
        - *Other countries internationally*

3.2 Mental Health

I have two hypothesis in the **mental health** family, the first, using mental health outcomes, while the second, using measures of digital addiction (Allcott et al., 2022).

3. **Changes in levels of subjective well-being.**

    - *On a scale of 1 to 5, indicate how much you agree with the following statements:*
        - *Over the past week, I was satisfied with my life*
        - *Over the past week, I was a very happy person*
    - *Please tell us the extent to which you agree or disagree with each of the following statements. Over the last week...*
        - *I was a happy person*
        - *I was satisfied with my life*
        - *I felt anxious*
        - *I felt depressed*
        - *I could concentrate on what I was doing*
        - *I was easily distracted*
        - *I slept well*

4. **Changes in levels of Digital Addiction.**

    - *On a scale of 1 to 5, how difficult is it for you to stop using SM once you log-in?*
    - *On a scale of 1 to 5, how satisfied were you with the time you spent on SM over the last one week?*
    - *How much time did you spend on the following platforms yesterday: Facebook, WhatsApp, Moj.*

3.3 Perception

I have one hypothesis in the **perception** family, which helps me understand the salience of the intervention as well as the strength of the first stage. These questions will only appear in the endline survey.

5. **Difference in exposure to customized content.**

   - *Do you agree (on a scale of 1 to 5) with the following statements about your SM in the last one week:*
     - *I saw the kind of content I wanted to see*
     - *I saw content that I usually do not see*
     - *I saw locally relevant content*
     - *I saw content that is relevant for my life*
     - *I saw content that was aligned with my worldview*
     - *I saw content about communities and people who are different from me*
   - Describe the amount of content you saw from each of the following categories as: (a) More, (b) Same, (c) Less
   - *Devotional*
   - *Nationalistic*
   - *News content*
   - *Political*
   - *Was something different about your SM feed in the past 6 weeks?* (Yes or No, choose one.) If yes, *what was it? We will call this different feed of the last few weeks your 'new' feed.* (free text answer.)

3.4 Behavioural Outcomes

I have two hypothesis in the **behavioral outcomes** family.

6. **Change in valuation of SM feed without the algorithm.** To establish user valuation of SM feed, and how that differs across treatment and control groups, I will offer respondents with a series of choices between keeping the 'new' feed vs. receiving cash prizes of different amounts in the endline survey (See Appendix B for the full schedule). One of these choices will be randomly selected as the choice that counts for some users chosen through a lucky draw. The valuation of the feed by users in the treatment group then gives us the demand for algorithms that customize users' content feed.

7. **Change in Network Externalities.** First, I gauge user types by offering users pairwise choices between content, without providing any other information. Then, I gauge whether users try to affect posts that other use see by 'gaming' engagement statistics on posts differentially by treatment and control groups. Both these questions are listed under the **Externalities** schedule in Appendix B. Other questions include:

   - *Do you think your liking behavior changed what other people saw on the platform?*

3.5 Demographics and Other Control Variables

The survey allows me to collect data for a set of demographic and other important control variables that I use to improve the precision of the estimates. This set of questions that enable data collections on these variables include:

- *Which of the following best describes your occupation:*
  - Self-employed in primary sector
  - Self-employed in secondary sector
  - Self-employed in tertiary sector (including shop-owner)
  - Salaried employee
  - Casual worker in agriculture
  - Casual worker in non-agricultural activities
  - Other
- *Are you a migrant worker? If yes, please state of district of origin.*

Data on other user characteristics is collected from the platform itself, these include: gender, age, region, date of account creation/ technology adoption, and mobile phone prices. In addition to user reported statistics, I will also infer user 'type' from their survey responses, as well as user engagement with content from different tag-genres, and a popular class of hash-tags in the baseline period. I will also measure 'hatefulness' of user by measuring their engagement with toxic and polarizing content in the baseline period using methods described in the previous section.

4 Empirical Analysis

4.1 Balance Checks

I will verify balance prior to the experiment across all treatment groups and the control group on the covariates. I will use experimental data from the platform for this analysis. I will report a balance table for the covariates, which will include the mean for the treatment group, differences relative to the control group, and results from the t-tests of the null hypothesis of zero difference. I will also regress the treatment variable on all the covariates simultaneously, and report the F-statistic for joint significance. In the absence of balance, I will match users on propensity scores constructed using observable characteristics following Abadie and Imbens (2016) and report estimated treatment effects from this procedure.

4.2 Identifying Assumptions

I assume that

$$(Y_i(0), Y_i(1)) \perp D_i | p(X_i)$$

where, $(Y_i(0), Y_i(1))$ are potential outcomes under binary treatment D_i, which is given by the second row in Table 1. That is to say, for $D_i = 1$, I populate user feed by picking posts uniformly at random from a corpus of 2 million posts, i.e. the degree of customization for the treatment group equals 0. Finally, $p(X_i)$ is the propensity score (Rosenbaum and Rubin, 1985) constructed using the list of observed user characteristics in Section 3.5.

4.3 Outcomes of Interest

I expect outcomes of interest to be influenced by the treatment assignment through aspects of platform *exposure*, including channels like:

$C_1 =$ hash tags of posts in content feed (hash-tag exposure)

$C_2 =$ hate speech in posts in content feed (hate speech exposure)

$C_3 =$ immoderacy and emotional appeal of content feed (extreme and emotional exposure)

$C_4 =$ spillovers in content exposure due to changed content feed for treated

I measure exposure right before the start of the experiment, in the last week of January 2023, and then six weeks after the intervention has been administered, up till April 2023.

Further, I divide the outcomes of interest into three categories: **(a)** On-platform, (b) Off-platform, (c) Neighbourhood Effects. There are various platform-based outcomes, like user engagement with content, that could be affected by the treatment, including:

$P_1 =$ tag-genre of posts in content feed (genre engagement)

$P_2 =$ hash tags of posts in content feed (hash-tag engagement)

$P_3 =$ hate speech in posts in content feed (hate speech engagement)

$P_4 =$ polarizing posts in content feed (engagement with politically polarizing content)

Off-platform outcomes are collected through the survey, and I test if the following are affected by the treatment:

$O_1 =$ subjective well-being and mental health

$O_2 =$ digital addiction

$O_3 =$ user perception about changes in content exposure

The possibility of stratification in treatment assignment at the neighborhood level would enable me to measure the effect of the intervention on political outcomes that are typically aggregated at sub-district levels. These include a mix of outcomes that are measured using the survey, and those that are obtained from publicly available data (merged with administrative data from SM with approximate user location captured by her neighborhood):

$N_1 =$ election results and winning party

$N_2 =$ protest eruption

$N_3 =$ political participation in institutions of local governance

$N_4 =$ increased offline interaction with peer group

4.4 Primary Outcome Treatment Effects

The general strategy to test each of the hypothesis laid out above is to regress each outcome of interest on a variable indicating treatment status as well as a set of controls. This gives the Intent to Treat (ITT) effects as outlined below. The experiment allows me to estimate the effect of assignment to the treatment group.

$$Y_i = \beta_0 + \beta_1 D_i + \gamma X_i + \varepsilon_i \tag{1}$$

Here, Y_i is the outcome of interest for individual i, D_i is the treatment dummy indicating whether a respondent is in the treatment group (with control group forming the base category in this instance), and X_i is a vector of individual level control variables. These covariates include information about on-platform characteristics and information collected in the survey during both the baseline period. The full list of variables is in Section 3.5.

4.5 Heterogeneous Effects

Lastly, I will examine treatment effect heterogeneity among various subgroups in the sample. I assume heterogeneous treatment effects across various sub-groups in the population, where sub-groups are determined by the list of observable characteristics in Section 3.5. These characteristics also include user ‘hatefulness’ at the baseline, as well as her affinity towards broad content genres, like Politics, Devotion, Music, etc. I rely on both standard regression based approaches, and machine learning techniques.

For the regression based approach, the modified ITT specification is presented below. For a sub-group defined by Z_i, β_3 is the coefficient of interest, which is the test for heterogeneity in treatment effects for the given subgroup. I will run this specification for each of the variables (both on and off platform variables) listed in 3.5.

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \beta_3 Z_i \cdot T_i + \gamma X_i + \varepsilon_i \tag{2}$$

The machine learning approaches use cross-validation to discover axes of heterogeneity. I follow two related approaches, firstly from Chernozhukov et al. (2018), and secondly, from Wager and Athey (2018) as well as Athey and Imbens (2016), toward this end. I will provide the full set of X_i from the survey and administrative data to let these unsupervised algorithms discover best ways to present heterogeneous effects in the data.

4.6 Non-parametric Methods

I will use non-parametric models, that do not require correctly specified functional forms, to provide robustness checks for the estimates uncovered using the parametric models. This is especially challenging, as it is important, due to large volume and dimensionality of the big data I am analyzing (Ng, 2017).

4.7 Corrections for Multiple Hypothesis Testing

Since I have multiple outcomes of interest within each hypothesis, and multiple hypothesis within each family of outcomes, I make the following adjustments to account for multiple hypothesis testing following Anderson (2008) and the pre-analysis plan for Björkegren et al. (2022).

- I construct an outcome index– the average of the z-scores of the outcomes, for each hypothesis. The index then serves as the main outcome of interest. Then, I repeat the analysis using an outcome index constructed using Principal Component Analysis.
- I will report the False Discovery Rate adjusted p-values for each individual outcome in a hypothesis.

References

A. Abadie and G. W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

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

M. L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495, 2008.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26 (10):1531–1542, 2015.

D. Björkegren, J. Blumenstock, O. Folajimi-Senjobi, J. Mauro, and S. R. Nair. Instant loans can lift subjective well-being: A randomized evaluation of digital credit in nigeria. *arXiv preprint arXiv:2202.13540*, 2022.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57 (3):535–74, 2019.

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.

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

A. Kalra. A'ghetto'of one's own: Communal violence, residential segregation and group education outcomes in india. 2021.

J. Kleinberg, S. Mullainathan, and M. Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776*, 2022.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

S. Ng. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. 2017.

T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social networks*, 35(2):159–167, 2013.

T. Piccardi, M. Gerlach, and R. West. Going down the rabbit hole: Characterizing the long tail of wikipedia reading sessions. *arXiv preprint arXiv:2203.06932*, 2022.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39 (1):33–38, 1985.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

D. J. Watts and S. H. Strogatz. Watts- 1998-collective dynamics of 'small-world. *Nature*, 393(6684):440–442, 1998.

Appendix A: Consent Form

Researchers: Aarushi Kalra

Brown University

+1-401-808-7950

RESEARCHER'S STATEMENT

We are contacting you to ask you to be in a research study. Your participation in this study is voluntary. The purpose of this statement is to give you information to help you decide whether to be in the study or not. We will state the purpose of the research, the possible risks and benefits, and your rights as a volunteer. When you have read these terms, you can decide if you want to be in the study or not.

PURPOSE OF THE STUDY

The purpose of the research is to better understand user preferences, and the factors that determine these preferences for SM users.

PROCEDURES

If you agree to participate, we will begin the survey, which will take 5 to 10 minutes. In the survey, we will ask you a series of questions about your general livelihood and well-being. We will not be collecting any additional personally identifying information beyond your name. As part of the study, we will link your survey responses to data already collected by SM. The data we collect will be shared with third party researchers and will be used for research purposes only. No one besides the principal investigators on this study will have access to these data, and any personal information will be removed as soon as this survey has been completed and entered into the computer. The survey is conducted in two stages, and we will send you a follow-up questionnaire within four weeks.

COMPENSATION

At the end of the survey, you will be compensated with Amazon gift vouchers worth Rs. 100. ***The full compensation amount will be sent to your amazon account only when you complete both stages of the survey.***

BENEFITS

Aside from the compensation, you will receive no direct benefit from the survey. Your responses can help us better understand how SM can cater to your content preferences.

RISKS

Some questions in the survey could make some people uncomfortable. For example, we may ask about your mental health. If you would prefer not to answer any individual question or group of questions, you can skip those questions. As with all research, there is a chance that confidentiality could be compromised; however, we are taking precautions to minimize this risk.

CONFIDENTIALITY

Your study data will be handled as confidentially as possible. No personally identifiable information will ever be shared with any third party.

SUBJECT'S RIGHTS If you agree to participate in this project, please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. The alternative is not to participate. You have the right to refuse to answer particular questions. If you agree to participate in this research study, please say so. If you have any questions or concerns about this study, you may contact Aarushi Kalra at +1-401-808-7950. If you have any questions or concerns about your rights and treatment as a research subject, you may contact SM at complianceofficer@SM.co.

Appendix B: Survey Instrument

Enclosed on next page.

Survey Instrument

Baseline

1. Attitudes

Now, I would like to learn your preferences over content. For the following pairs of posts, please select only one option.

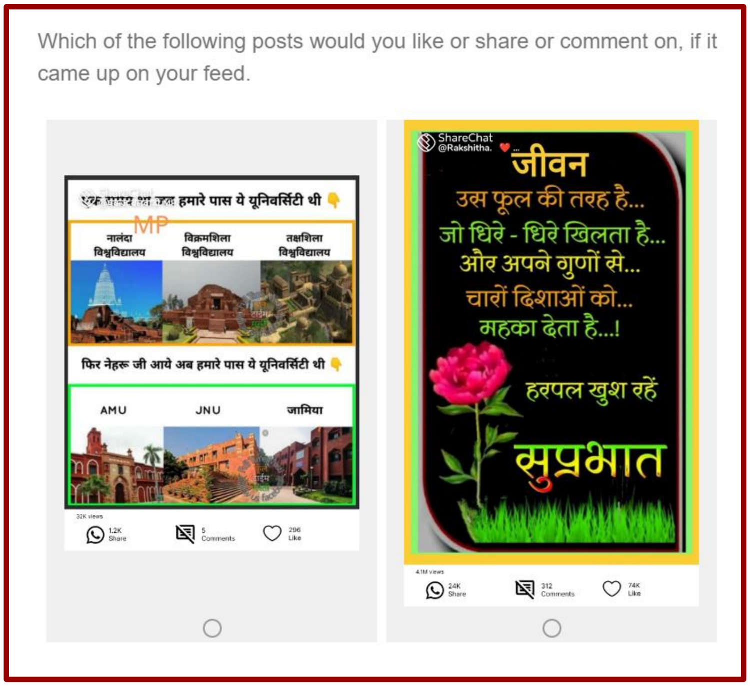
a. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. Pro-BJP post without engagement stats
   ii. Cute cat video without engagement stats

b. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. anti-BJP post without engagement stats
   ii. Cute cat video without engagement stats

c. Choose the physical radius you would like to see news on SM from (select multiple):
   i. Your neighborhood
   ii. Your city
   iii. Your state
   iv. All India
   v. USA
   vi. Other countries Internationally

2. Willingness to Pay

Researchers at Brown University in the USA are developing an app, called PersonalizeMyFeed, that will help improve personalization of your SM feed. This app works by collecting more information about your content preferences by asking you some questions if you consent to answer them. The decision to get the app is entirely yours. Below we show you an example of the kind of questions we will ask you on the homepage of PersonalizeMyFeed app.


Which of the following posts would you like or share or comment on, if it came up on your feed.

To reiterate, your decision to get (or not) get this app does the following:

Get the PersonalizeMyFeed App	Answer some questions to help us learn your preferences over content
Do not get the App	Do not answer these questions and keep your SM feed the same

To establish your valuation of this new app, we will offer you a series of choices between downloading PersonalizeMyFeed app vs. receiving cash prizes of different amounts. One of your choices will be randomly selected as the choice that counts for some users chosen through a lucky draw.

1. Which of the following do you prefer?
    a. Get PersonalizeMyFeed App
    b. Do not download the app BUT you receive Rs. 600
2. Which of the following do you prefer?
    a. Get PersonalizeMyFeed App
    b. Do not download the app BUT you receive Rs. 400
3. Which of the following do you prefer?
    a. Get PersonalizeMyFeed App
    b. Do not download the app BUT you receive Rs. 200
4. Which of the following do you prefer?
    a. Get PersonalizeMyFeed App
    b. Do not download the app BUT you receive Rs. 150
5. Which of the following do you prefer?
    a. Get PersonalizeMyFeed App
    b. Do not download the app BUT you receive Rs. 100
6. Which of the following do you prefer?

   a. Get PersonalizeMyFeed App
   b. Do not download the app BUT you receive Rs. 50

7. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App
   b. Do not download the app BUT you receive Rs. 0
8. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 50
   b. Do not download the App
9. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 100
   b. Do not download the App
10. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 150
   b. Do not download the App
11. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 200
   b. Do not download the App
12. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 400
   b. Do not download the App
13. Which of the following do you prefer?
   a. Get PersonalizeMyFeed App AND receive Rs. 600
   b. Do not download the App

4. Demographics

Finally, to better understand your preferences for content, I will ask you for some general information about yourself. Please answer the following.

   a. In which district do you currently reside? (free text)
   b. Which of the following best describes your occupation:
      i. Self-employed in primary sector
      ii. Self-employed in secondary sector
      iii. Self-employed in tertiary sector (including shop-owner)
      iv. Salaried employee
      v. Casual worker in agriculture
      vi. Casual worker in non-agricultural activities
      vii. Other
   c. Are you a migrant?
      i. Yes
      ii. No
   d. If yes, state district of origin?

Endline

1. Measures of Well-being

Now I will ask you questions about your general well being and state of mind in the last week.

a. On a scale of 1 to 5, indicate how much you agree with the following statements:
   - i. Over the past week, I was satisfied with my life
   - ii. Over the past week, I was a very happy person

b. On a scale of 1 to 5, how difficult is it for you to stop using SM once you log-in?

c. On a scale of 1 to 5, how satisfied were you with the time you spent on SM?

d. How much time did you spend on the following platforms yesterday? Please use the slider to capture the number of minutes spent.
   - i. Facebook
   - ii. WhatsApp
   - iii. Moj
   - iv. In-person meeting with neighbors

e. Please tell us the extent to which you agree or disagree with each of the following statements. Over the last week…

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I was a happy person I was satisfied with my life I felt anxious I felt depressed I could concentrate on what I was doing I was easily distracted I slept well					

2. Attitudes and Externalities

a. Choose the physical radius you would like to see news on SM from (select multiple):
   - i. Your *basti*
   - ii. Your city
   - iii. Your state
   - iv. All India
   - v. USA
   - vi. Other countries Internationally

Now, I would like to learn your preferences over content. For the following pairs of posts, please select only one option.

b. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. Pro-BJP post without engagement stats
   ii. Cute cat video without engagement stats
c. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. anti-BJP post without engagement stats
   ii. Cute cat video without engagement stats

Now, I am asking you to choose between the same set of posts as above, but would like you to also pay attention to the REAL engagement statistics depicted below the posts, that I had previously not shown to you.

d. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. Pro-BJP post with less likes
   ii. Cute cat video with more likes
e. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. Anti-BJP post with more likes
   ii. Cute cat video with less likes
f. Which of the following posts would you like or share or comment on, if it came up on your feed.
   i. Pro-BJP post with more likes
   ii. Cute cat video with less likes
g. Which of the following posts would you like or share or comment on, if it came up on your feed
   i. News about your Municipality or your *Panchayat*
   ii. News about your state's Chief Minister

3. Strength of First Stage

SM is continuously trying to improve the customer experience. So I will now ask you some questions to understand if you have noticed any changes on your app during the previous week.

a. Do you agree (on a scale of 1 to 5) with the following statements about your SM feed in the last one week:
   i. I saw the kind of content I wanted to see on SM
   ii. I saw more viral content
   iii. I saw more content that I agree with
   iv. I saw more content on SM from my district
   v. I saw more content that surprised me
   vi. I saw more content that I discussed with my friends
   vii. I saw content that is relevant for my life
   viii. I saw more content that made me angry

        ix. I saw content that was aligned with my worldview
        x. I saw content about communities and people who are different from me
    b. Select one of three options about the amount of content you saw from the following categories: (a) I saw more of such content (b) I saw less of such content (c) Same
        i. Devotional
        ii. Nationalistic
        iii. Locally Relevant
        iv. News
        v. Political
    c. Did you think your liking behavior changed what other people saw on SM?
        i. Yes
        ii. No
    d. Was something different about your SM feed in the past 6 weeks?
        i. Yes
        ii. No
        iii. Can’t say
    e. If Yes, what was it? We will call this different feed of the last few weeks your ‘new’ feed. (optional text question - conditional on previous q response)

4. Willingness to Pay

You may have noticed some changes to your SM feed recently. Based on your experience with SM in the last one week, we would like to implement ways to improve your experience and evaluate changes that SM has made to your feed in the last one month. Let us call the recent changes in your SM feed as your ‘new’ feed. To establish your valuation of this new feed, we will offer you a series of choices between keeping the new feed for the next one year vs. receiving cash prizes of different amounts. One of your choices will be randomly selected as the choice that counts for some users chosen through a lucky draw.

1. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 600
2. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 400
3. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 200
4. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 150
5. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 100

6. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 50
7. Which of the following do you prefer?
    a. Keep the new feed for the next one year
    b. Revert back to the old feed BUT you receive Rs. 0
8. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 600
    b. Revert back to the old feed
9. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 400
    b. Revert back to the old feed
10. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 200
    b. Revert back to the old feed
11. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 150
    b. Revert back to the old feed
12. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 100
    b. Revert back to the old feed
13. Which of the following do you prefer?
    a. Keep the new feed for the next one year AND receive Rs. 50
    b. Revert back to the old feed

6. Demographics

Finally, to better understand your preferences for content, I will ask you for some general information about yourself. Please answer the following.

e. In which district do you currently reside?
f. Which of the following best describes your occupation:
    i. Self-employed in primary sector
    ii. Self-employed in secondary sector
    iii. Self-employed in tertiary sector (including shop-owner)
    iv. Salaried employee
    v. Casual worker in agriculture
    vi. Casual worker in non-agricultural activities
    vii. Other