Social Media, Discrimination, and Politics: Online Field Experiments in India

Jun Goto*

12 September, 2019

Abstract

This paper provides field-experimental evidence on how repeated exposure to biased news through social media affects societal discrimination and political polarization in India. I developed and released an original mobile application in which social media news is delivered to users' mobile phones on the daily basis during the experimental period (202 days). In this application, treatment and control arms are randomly built-in. In the treatment arm, a machine learning algorithm customizes news circulation based on individual user's past behavior on the web. In the control arm, news circulation is fully randomized. Three main findings are obtained by econometric analyses. First, the customized news circulation significantly makes an individual flow of daily news more religiously biased and politically extreme. Second, users' beliefs and preferences are unlikely to be affected by repeated exposure to such slant news. However, if users can recognize that a majority of other users express positive opinions on the slant news, they get to be less against religious discrimination and more strongly favor political extremism. Finally, the cognitive bias rather than concerns for social reputation is a main driver of emergence of societal discrimination and political polarization.

*Goto: Kobe University, 1-1 Rokkodaicho, Nada Ward, Kobe, Hyogo 657-0013, Japan, goto-jun.jg@gmail.com

1 Introduction

The Internet and social media services penetrate deeply into our society. Not only developed countries but also developing countries experience the rise of social media. In fact, according to Pew Research Center, the intensity of usage of the Internet and social networking sites have sharply grown in developing countries during this decade, and they are rapidly catching up with developed countries: the ratio of people who use the Internet at least occasionally has risen from 42% in 2013-14 to 64% in 2017-18 in developing countries while the ratio is relatively stable in developed countries: 86% in 2015-16 and 87% in 2017-18. This filling-gap-trend is more prominent in usage of online networking sites: the share of people who use social networking sites has increased in developing countries from 34% in 2013-14 to 53% in 2017-18 which is very close to the level of developed countries, i.e., 60% in 2017-18.

In this situation in mind, global experts and many politicians express concerns about expected consequences induced by this rapid prevalence of social media services. Sunstein (2018) claims that as the Internet grows more sophisticated, it causes new threats to democracy. Social media can sort us ever more efficiently into groups of the like-minded, creating *echo chambers* that amplify and reinforce our views through repeated communication within a closed platform. This might result in the situation that people of different political views cannot even understand each other. Pariser (2011) also warns the public of negative consequences of personalized web search algorithms which can cause *a filter bubble* in which people are isolated each other based on their own cultural or ideological bubbles and become separated from information that disagrees with their viewpoints. Despite its importance on every aspect of our life, however, the influence of social media on our social and political preferences are still under-investigated due to data limitation and a lack of exogenous variations.

This paper fills this gap by providing field-experimental evidence on the following research questions: does the rapid increase of social media usages as a news source deteriorate our tolerant societies? To be more precise, does social media (or AI-customized news circulation) facilitate users to receive more "biased" news which is relatively extreme on the side of own ideology, compared to neutral news contents? If so, does more frequent exposure to "biased" social media news exacerbate the degree of discrimination and political polarization in a society?

In order to answer above questions, I developed an original mobile application in which news articles on politics in India are distributed to users on the daily basis. The treatment

and control arms are built-in this application: in the control group, the mobile application randomizes news articles which are circulated into a user's mobile phone every day. In the treatment group, the simple AI algorithm customizes news circulation for each user based on his or her history of web browsing behavior. Within a treatment/control group, I provide additional randomizations: first, the number of news per day is randomized between 1 to 10. Second, an opportunity that a user can check others' opinions on the news which is delivered to his or her smart phone is also randomized. These exogenous variations allow me to identify whether the AI-customized news circulation induces distorted distribution of news consumption by users, and how other users' responses on biased news affect evolution of users' preferences toward politics and religious discrimination.

To explicitly measure the degree of discrimination and political extremism, lab-in-the-field experiments are also set up inside program source codes of the mobile application. Other-other allocation games which are modified versions of Chen and Li (2009) are involved in the start-line and end-line surveys, which are aimed at measuring religious-based discrimination and political preferences.

I released β-version of this application to those who live in the South India in December 2015. I target the population who actively use SNS services such as Twitter and/or Facebook. In total, 25,966 users have accepted to cooperate on my research project. They had been exposed by a flow of news which is circulated through the mobile application for 202 days. My experimental design has four advantages and novelty: first, this is the first field experiment which is well-designed to answer the global concerns on the prevalence of social media, i.e., is social media a echo chamber and a filter bubble, which results in further riving our society? To test this hypothesis, the random assignment of news curation algorithms, namely (a) the randomized news curation arm and (b) the AI-customized algorithm arm, is necessary since it enables me to precisely quantify the causality of a recent curation mechanism on the distribution of news articles' bias that a user experiences during the experimental period. Moreover, exclusively focusing on the sample in (a) the randomized news curation arm, I can exploit exogenous variations on the degree of slant of each news article to answer whether repeated exposure to such slant news really facilitates societal discrimination and political polarization. Second, in order to scrutinize the dynamics of religious discrimination and political polarization, the short-term intervention might not be sufficient to observe the changes of user's behavior and belief so the long-term intervention is required. My experimental design obtains the relatively long-term panel data on the individual user level and thus overcomes this obstacle. Third, the randomized display of other users' opinions is critically important when

we replicate "real" online communities and news consumption. In social media, almost no news articles flow into a user's device without others' evaluation or responses. A comparison of the impacts of news between with and without others' responses approximately corresponds to the different structures of information circulation between social and mass media. Therefore, this design is indispensable to be incorporated to derive practical and feasible implication for improving the design of the application for social media news in general. Finally, the intervention through the original mobile application allows me to extract online social network information as well as other web browsing behavior on a user's smart phone. This information can be utilized to analyze heterogeneity impacts of biased news in terms of a scale of social networks and further investigation on the mechanisms behind the causal link between social media and discrimination.

I obtain three main results. First, the customized news circulation significantly makes an individual flow of daily news more religiously biased and politically extreme than the randomized news curation: the more likely users are to favor Hindu nationalism at the beginning of the intervention, the more distorted the distribution of news bias for them toward Hindu nationalism and vice versa. Second, users seem not to be affected by repeated exposure to such slant news. Interestingly, however, if users can recognize that a majority of other users express positive opinion on the slant news, they get to be less against religious discrimination and more strongly favor political extremism. Finally, I specify the underlying mechanism behind this belief shift. In my context, the cognitive bias i.e., confirmation bias, rather than the concern for social reputation is a main driver of emergence of societal discrimination and political polarization: users arbitrarily pick up news articles and interpret that they reflect reality only if those articles affirm their ideology and other users support them.

This paper contributes four strands of literature. First, many scholars in social sciences shed light on the impacts of media on social, economic, and political behavior. Adena, Enikolopov, Petrova, Santarosa, and Zhuravskaya (2015) examines the impact of radio on the spread of political extremism or the rise of Nazi. Gerber, Karlan, and Bergan (2009) conduct a field experiment to measure the effect of exposure to newspapers on political behavior and opinion, focusing on the 2005 Virginia gubernatorial election. Gentzkow (2006) uses variation across markets in the timing of television's introduction to identify its impact on voter turnout. Enikolopov, Petrova, and Zhuravskaya (2011) quantifies the causal impact of national TV channel independent from the government on voting behavior and political preferences. Enikolopov, Makarin, and Petrova (2017) show that penetration of VK, the dominant Russian online social network, leads to more

protest activity during a wave of protests in Russia in 2011. I complement this literature by providing first experimental evidence of the impacts of AI-customized news circulation, repeated exposure to news slant, and communication with like-minded people on societal discrimination and political extremism. Second, this paper is closely related to the literature on an origin, a spread, and evolution of nationalism. Few studies are accumulated in this field but an exception is DellaVigna, Enikolopov, Mironova, Petrova, and Zhuravskaya (2014). They exploit a unique setting where cross-border nationalistic Serbian radio is available in a part of Croatia after WWII. They show that the exposure to this media triggers ethnic hatred toward Serbs in Croatia. I complement this literature by adding new evidence on the emergence of nationalism and its mechanism: why people get inclined to favor nationalism and what nurtures out-group hostility. Third, recently, we have seen that an increase of researches employing field experimental methods to identify the determinants of discrimination. Although a growing body of researches focus on this area as reviewed in Bertrand and Duflo (2017), how and under what conditions discrimination evolves is still largely unknown. Our evidence can fill this gap. Finally, evidence that my econometric specifications reveal corroborates theories of herding behavior (Banerjee, 1992) and conformity (Bernheim, 1994).

The rest of the paper is organized as follows: Section 2 explains political background of the targeted field, or the state of Kerala in South India. Section 3 elaborates experimental design, followed by data description in section 4. Section 5 shows estimation strategies and their results. Finally, I conclude in Section 6.

2 Background

Field experiments target the entire population of Kerala who uses SNS services at the beginning of intervention. Focusing on this region, this section explains the rise of Hindu nationalism and subsequent proliferation of discrimination. Then, I proceed to provide brief explanation on the background context of politics in Kerala.

2.1 Hindu Nationalism and Discrimination

Recently, as its economy has opened to the world and grown really robustly, Hindu nationalism has arisen and expanded to a whole of India. Hindu nationalism is the idea that the Hindu faith and culture should shape the state and its policies. One of the most typical example of Hindu nationalists is Narendra Modi, India's prime minister who was

elected in 2014. Since that election, Mr. Modi has defined himself as the representative in a conservative Hindu political movement that strives to make India a Hindu state. Furthermore, the Bharatiya Janata Party, or BJP to which Mr. Modi belongs, has been spreading an us-versus-them philosophy in the country. Consequently, the Hindu right has been more enfranchised at every level of government than before. This rise of nationalism might deteriorate a tolerant society founded by religious- and caste-based diversity in the country and New York Times explicitly expressed the concern of this phenomenon:[1]

> The consensus among Indian activists and liberal political analysts is that their society, under Mr. Modi, has become more toxically divided between Hindus and Muslims, between upper and lower castes, between men and women.

Kerala is not an exception of this nationwide movement. Historically, Kerala has been characterized as the most tolerant society in terms of a caste system and religion other than Hindu. However, the violence toward non-Hindu and political riots to ostracize other religion from this state have surged during this decade, as the national movement has collected growing attention. According to the official report published by National Crime Records Bureau (NCRB), *Crime in India 2012 Statistics*, Kerala is the place which experiences the largest number of riots in 2012: the number of riots in 2012 in Kerala becomes thirteen times as large as the number in Delhi. Concurrently, many middle- and upper-caste Hindus in Kerala resent longstanding affirmative action policies to help lower castes, and the special customary laws that allow India's Muslims to follow Islamic traditions when it comes to family legal matters like divorce and inheritance.

2.2 Politics in Kerala

The Kerala state Legislature is unicameral and has a membership of 141, where 140 are elected and one is nominated from the Anglo-Indian community. Elections are held to choose representatives to the civic bodies at various levels within the State, and the state has consistently come out with a voter turnout of 70% or above in almost all elections which have ever been held.

There are two major political coalitions in Kerala. The United Democratic Front (UDF) is the coalition of centrist and centre-left parties led by the Indian National Congress. The Left Democratic Front (LDF) is the coalition of left wing and far-left

[1] *Under Modi, a Hindu Nationalist Surge Has Further Divided India*, 11 April 2019: https://www.nytimes.com/2019/04/11/world/asia/modi-india-elections.html

parties, led by the Communist Party of India (Marxist). In terms of Hindu nationalism, while UDF and LDF stand in a modest position, National Democratic Alliance (NDA) which consists of BJP and other right-wing parties radically advocates this ideology.

The last Kerala legislative assembly election, which was conducted in May 2016, had two contested grounds: do voters punish the incumbent government because local mass media revealed the government has committed corruption in every level of local governments? More importantly in this paper's context, the election also asked voters to answer how much they prefer to reflect Hindu nationalism in a policy making. The result of this election is as follows: voter turnout was 77.53%, up from 75.12% in the previous election. LDF won the election by securing 91 seats, defeating the incumbent UDF, led by the Indian National Congress (INC), which could only win 47 seats in the election. NDA which consists of BJP won only two seats. Although BJP which advocates Hindu nationalism lost the election, the share of the vote was 10.6% which was the third-rank followed by Communist Party of India (Marxist) and INC. This implies that the non-negligible number of voters have actually supported Hindu nationalism in the last election.

3 An Experimental Design

This section describes the design of two types of experiments: the social media experiment which is online field experiment using the mobile application, and lab-in-the-field experiments.

3.1 The Social Media Experiment

3.1.1 Basic Features of the Mobile Application

In order to investigate the impact of repeated exposure to slant social media news, I developed an original mobile application, *NowCast*. NowCast is the simple application for news curation with two features. First, the database of this application is massive, comprehensive, and instantaneous. The database that I rely on is operated by the *GDELT Project*: GDELT monitors print, broadcast, and web news media in over 100 languages from across every country in the world to keep continually updated on breaking developments anywhere on the planet. Anyone who is familiar with SQL can easily access to the database through the Google Cloud Platform. Its historical archives stretch back to January 1, 1979 and update every 15 minutes. I build computer programs to automatically extract *all* of news articles every day (at 7 am) which are categorized as either "politics

in India" or "social issues in India" and published within the last 24 hours. An average number of news articles in each day that I collect during the experimental period is 254.3. Within this sample articles, the application randomly delivers a certain number of news between one to ten to users' devices every day. Second, I use machine learning techniques to create a summary of each news. Therefore, a user doesn't need to visit news source URLs to learn main stories of news articles, which can save his or her time. This can be an advantage of my application in comparison with other existing news curation services and attract potential users who are not interested in previous news curation apps.

Other common features and designs of news curation services are also adopted in this app: evaluation buttons (like and dislike such as in Facebook) on each news, a comment box for own opinion, a display of other users' comments, a sharing button through major SNS services like Twitter and Facebook.

3.1.2 A Design of Randomization

Treatment and control arms are randomly involved in this application. In the treatment arm, a customized-news-circulation-algorithm is built in the program source code. Thus, machine learning techniques automatically calculate an optimized news flow for each user based on his or her clicking behavior on all of websites and browsing behavior on my application. In the control arm, a news circulation is fully randomized.

In each treatment/control arm, I incorporate sub-treatments: first, a quantity of news per day is randomized between one to ten; second, other users' comments randomly show up on a user's interface, which is executed by a news level rather than an individual user level. For example, if a user receives four news, two of them may come with others' comments while the rest of them come without comments. By contrast, a user is always permitted to write down his or her own opinion on each news.

The randomized display of other users' comments in each news needs more explanation. This randomization takes two steps. In the first step, the app creates three distinct databases of users' comments in every news. The first database consists of positive comments only in which users express agreement to a news content or its tone. And the second database consists of negative comments only in which they express disagreement. The third database consists of neutral stance comments only. Note that this database is formulated in a unit of each individual news. As mentioned above, the app collects candidates of news to be circulated from the GDELT database at 7 am every day and immediately broadcasts them to users. All of users can start to write down their opinions right after the circulation. The app accumulates their comments until 3 pm on that day.

At 3 pm, the app is programmed to allocate all of comments in each news to above three databases using a computational text analysis method. So, how does the app categorize comments into three different dispositions? I built in a unique question for this task. A comment box says "what do you think of this article? True or false, agree or disagree, and biased or not biased. Post your opinion." This question allows me to evaluate comments in a single dimension, i.e., positive vs. negative. Additionally, this question assures that comments refer to news article rather than a general opinion. Then I apply one of the machine learning methods which is called wordfish to estimate a position of comments on a positive-or-negative dimension.[2]

In the second step, the app determines how many positive, negative, and neutral comments are displayed in a user's screen. Suppose that a user receives a certain amount of news which is equal to $N_{news,d}$ in day d. Then the app draws the first lottery $N_{news,d}$ times and allocates a set of news into two categories: a display with or without others' comments. The latter category offers a news content only (without comments). When the news is assigned to the former category, it always comes with three comments. In order to select three comments from three databases explained above, the app draws the second lottery three times for that news. This second lottery can take three integers from one to three. The number one corresponds to a positive comment; the number two corresponds to a neutral comment; the number three corresponds to a negative comment. so, there are potentially nine ($= 3 \times 3 \times 3$) results. Note that if a user is allowed to see others' comments based on the randomized assignment, a total number of comments is always three and only the distribution of positive, neutral, and negative comments is exogenously different.

3.2 Lab-in-the-field Experiments

To elicit the degree of caste-based and religious-based discrimination, I conducted an other-other allocation game which is the modified version of Chen and Li (2009). In this experimental game, user i is matched with two other users, j and k. No information on j and k is revealed to user i except for their caste and religion information. The experimenter gives user i an endowment 500 Rs and asks him or her to allocate *all* of an endowment to user j and/or k. Any amount of transfer can be sent to j and k given an amount of an endowment, and no money can be left for user i. Importantly, I set user

[2]Wordfish is a Poisson scaling model of one-dimensional document positions (Slapin & Proksch, 2008). It is an unsupervised one-dimensional text scaling method, meaning that it estimates the positions of documents solely based on the observed word frequencies.

j's caste and religion is always same with user i: if user i's caste is *Nair* so is user j; if user i's religion is Christianity rather than Hindu so is user j. Furthermore, I exogenously alter user k's religion and caste from top to bottom in the religious hierarchy. To be more precise, there are three religion, i.e., Hindu, Islam, and Christianity, and twelve castes within Hindu. So a total number of religious categories is fourteen. Then user i is asked to play an other-other allocation game fourteen times where a user j's religious category is always fixed while a user k's religious category varies from the first to the fourteenth.

I construct outcome variables regarding the degree of discrimination based on a bunch of other-other allocation games in a following way:

$$Outgroup\ discrimination = \sum_{k}^{n_{all}} (transfer_{ij}^{k} - transfer_{ik})/n_{all}. \tag{1}$$

This first outcome variable is an average outgroup discrimination where $transfer_{ij}^{k}$ denotes an amount of transfer by user i to user j conditional on matching with user k; $transfer_{ik}$ denotes an amount of transfer by user i to user k; n_{all} denotes a total number of religious categories which is equivalent to fourteen in this case. Additionally, I can also define an average outgroup discrimination against only *upper* castes compared to oneself: $\sum_{k}^{n_{upper}} (transfer_{ij}^{k} - transfer_{ik})/n_{upper}$ where I calculate outgroup discrimination against only upper castes; a total number of upper castes compared to oneself is n_{upper}. I can obtain an outcome variable against only lower castes by the symmetric calculation as well.

In addition to this experiment, I also conducted a third-party punishment game developed by Fehr and Fischbacher (2004). In this game, user m with the same caste of user i is matched with user h who belongs to other castes or religion. user i is asked to play as a third-party punisher in the standard dictator game between user m and h (user m is a receiver and user h is a sender). User i can punish user h if she is not satisfied with a sending amount by user h. The minimum amount of transfer that user i can accepts measures his or her spitefulness to user h.[3] As in an other-other allocation game, user i is asked to play this game fourteen times in which a user h's social class is varied. I define that user i has spiteful discrimination to user h if user i always commits punishment regardless of a user h's sending amount to user m in the dictator game.

Finally, I implemented a modified version of an above other-other allocation game to identify user's political preferences. In this game, first the experimenter gives 500 Rs

[3]I employed the strategy method to identify the minimum acceptance amount of transfer by asking the following question: how much transfer at least do you (user i) require user h if you should give up your punishment option toward him or her?

as an endowment to user i. Then he or she is asked to allocate all of that endowment to political candidates as donation. Political candidates here are actual candidates of his or her electoral constituency in the Kerala 2016 legislative election. This game was conducted right before the election and the transfer from user i to each political candidate was realized right after the election so that the experimental donation couldn't affect electoral results directly.

Timing, Procedures, and Target of Payments. Other-other allocation games and third-party punishment games were conducted through the mobile application that I developed. The baseline lab-in-the-field experiments were implemented right after the download of the app (during a week from December 11, 2015 to December 17, 2015), while the endline ones were done right before election (during a week from April 24, 2016 to April 30, 2016). Rewards earned by users in lab-in-the-field experiments can be exchanged with equivalent e-commerce currency.[4] However, there are some requirements to realize rewards: in the app, a user can accumulate the original token called NowCast Point. He or she can earn one NowCast Point (1) if he or she reads all articles within 24 hours that are newly circulated to his or her smart phone; (2) if he or she writes down an opinion in a comment box. Every time an individual total NowCast Point increases by thirty, the app provides one-seventh of total rewards with a user. If a user can earn more than 240 points by the 2016 election, I give him the qualification to receive all of rewards that would be obtained in the endline lab-in-the-field experiments. After surpassing 240 points, the app exchanges NowCast Points with equivalent e-commerce currency by ten points.

This dynamic and non-linear incentive scheme sufficiently attracts users, keeps them active for a long-term, and prevents them from uninstalling the app.[5]

I target the population who lives in Kerala, above 18 years old as of the intervention, and can be accessible through SNS services. The size of population is more than 8 million and the total number of those who downloaded the app is 34,525. only after receiving approval to cooperate this research project by users, I get authorized to utilize all of behavioral data. The final size of observation is 25,966.

[4] A user can select with which e-commerce currency he or she wants to exchange among several options such as Amazon and Flipkart etc.

[5] As a result, even after a whole of experiment periods or 202 days, an attrition rate is quite low in my sample, that is, 0.12%.

4 Text Data and Hindu Nationalism

This section provides the detailed explanation on how to construct a text-based slant index for news articles. I define that a news article is slant to Hindu nationalism if text information extracted from that news article is structurally similar to statements made by national congress men or politicians who define themselves Hindu nationalists. By contrast, a news article is slant to anti-Hindu nationalism if text information is similar to anti-Hindu nationalists who are also national congress men, explicitly against Hindu nationalism by seeking religious diversification and a tolerant society.

4.1 Preparation for Computational Text Analyses

First, I combine all of Hindu nationalists' statements into the single corpus, using the minutes records regarding all of national congress meetings held in 2015. Similarly, I also create the different corpus composing only statements by anti-Hindu nationalists in the same national congress meetings. [6] I identify whether each politician or national congress man is a Hindu nationalist, an anti-Hindu nationalist, or a person who stands with a neutral stance toward Hindu nationalism by investigating his or her political preference extrapolated from manifestos and partisanship.

Second, I construct *a word-document matrix* from all of news articles that were circulated to at least one user during a whole of an experimental period plus above two corpora consisting of Hindu nationalists' statements and anti-Hindu nationalists' statements in 2015. In this matrix, the number of rows is equal to the number of unique words shown in either news articles or two politician's corpora, and the number of columns is equal to the number of unique documents which are unique news articles that at least one user received during the experimental period plus two politician's corpora. In my context, the number of unique words is 29,382 and the number of unique documents is 38,298 in which 38,296 corresponds to the number of unique news articles. Note that in order to make variables created by a computational text mining plausible, I remove stopwords from text information such as prepositions ("of" and "on") and articles ("the" and "an").

[6] A *corpus* is a dataset which is composed by texts or politicians' statements in this case.

4.2 Computational Text Analyses

A Machine Learning Method. I apply a machine learning technique to estimate consistency or similarity between each news article and the (anti-) Hindu nationalist corpus. More specifically, I adopt latent Dirichlet allocation (LDA). LDA is an unsupervised generative model that assigns topic distributions to documents. At a high level, the model assumes that each document will contain several topics, so that there is topic overlap within a document. The words in each document contribute to these topics. The topics may not be known a priori, and need not even be specified, but the number of topics must be specified a priori. Note that there can be words overlap between topics, so several topics may share the same words. The model generates to latent variables: a distribution over topics for each document and a distribution over words for each topics. After estimation, each document will have a discrete distribution over all topics, and each topic will have a discrete distribution over all words. After having a topic distribution for all documents, I turn to calculate statistical similarity between documents. To do so, I compare the topic distribution of a certain document of news articles to the topic distributions of a document either Hindu nationalist corpus or anti-Hindu nationalist corpus. I use the Jensen-Shannon distance metric to estimate the similarity between them. First, I need to calculate the Jensen–Shannon divergence which is a method of measuring the similarity between two probability distributions. For discrete distirbutions P and Q, the Jensen-Shannon divergence, JSD is defined as

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{2}$$

where $M = 1/2(P + Q)$ and D is the Kullback-Leibler divergence:

$$D(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$
$$\Rightarrow JSD(P||Q) = \tfrac{1}{2}\sum_i \left[P(i) \log\left(\frac{P(i)}{\frac{1}{2}(P(i)+Q(i))}\right) + Q(i) \log\left(\frac{Q(i)}{\frac{1}{2}(P(i)+Q(i))}\right)\right]. \tag{3}$$

The square root of the Jensen-Shannon divergence is the Jensen-Shannon Distance. The smaller the Jensen-Shannon Distance, the more similar two distributions are and in our case, the more similar any two documents are. Therefore, I assume that this Jensen-Shannon distance reflects the degree of slant toward (anti-) Hindu nationalism on each news article since it measures how similar a given news article is with (anti-) Hindu nationalist corpus.

Graphical Checks for Validity of Variables. Here, I present a graphical result of machine learning estimation on the consistency among documents to prove validity of each

variable. Figure 1 illustrates a two-way plot on the similarity indices with Hindu nationalists and with anti-Hindu nationalists. First, each variable has a plausible variation with positively-skewed distribution. This implies that, basically speaking, news articles are not so biased toward neither Hindu nationalism nor anti-Hindu nationalism. Second, two similarity indices are in inverse proportion. This means that machine learning techniques properly estimate the degree of slant on each article.

Non-machine Learning Methods. Alternatively, I construct another index for the degree of news bias, which is based on a simple count method rather than machine learning techniques. First, I decompose a news article into a list of words. Second, I pick up each word from the top of list and check how many times (anti-) Hindu nationalists use that word in 2015 congress meetings. Also, I calculate how many times all congress men use that word in the same meetings. Third, the ratio of the frequency that (anti-) Hindu nationalists use that word to the frequency that all politicians use is defined as a slant index for that specific word. Finally, I add all slant indices for each word shown in the list. This is the slant index for the targeted article.

5 Econometric Analyses

5.1 Randomized vs. Customized News Circulation.

First, I investigate the impacts of the treatment of AI-customized news circulation, compared to randomized news circulation, on a set of slant indices. To do so, I estimate the following simple equations:

$$slant_{ij} = \alpha_j + \gamma customized_{ij} + \eta X'_{ij} + \omega_{ij}, \tag{4}$$

$$antislant_{ij} = \alpha_j^{anti} + \gamma^{anti} customized_{ij} + \eta^{anti} X'_{ij} + \omega_{ij}^{anti}, \tag{5}$$

where $slant_{ij}$ and $antislant_{ij}$ denote average slant indices toward Hindu and anti-Hindu nationalism for user i in district j during the experimental period; $customized_{ij}$ denotes the treatment dummy which takes one if user i is assigned to the AI-customized news circulation group and zero otherwise; α_j and α_j^{anti} are district fixed effects; X_{ij} is a vector of control variables. In addition to an average slant index, I also examine a standard deviation of a slant index because the treatment might shrink a variation of news bias within an individual user: for example, the more they favor Hindu nationalism, the more they consume biased news toward Hindu nationalism and the less they consume articles which are distant from Hindu nationalists' statements.

Table 1 presents main impacts of customized news circulation on slant indices estimated by machine learning techniques. Panel A shows that the results for slant indices in terms of Hindu nationalism while Panel B shows the results for slant indices in terms of anti-Hindu nationalism. Column (1) demonstrates that the AI-customized news circulation (AI-CNC) doesn't have any influence on an average degree of bias toward Hindu nationalism with regard to news consumption by users. Column (5) also represents that the same result with Column (1) holds for bias toward anti-Hindu nationalism. Further analyses reveal what happens behind these results: Column (3) and Column (4) demonstrate that the treatment of AI-CNC expands the range of a slat index in the full sample because coefficients on the maximum and the minimum values are positive and negative with statistical significance, respectively. The same inference holds for a slant toward anti-Hindu nationalism by looking at Column (7) and (8). Importantly, according to column (2) and (6), the treatment decreases standard deviation of slant indices within individual users and thereby I can conclude that each user consumes a narrower range of news contents in terms of bias toward both Hindu and anti-Hindu nationalism. I implement robustness checks for variables created by machine learning techniques. Table 2 illustrates that the estimated results in Table 1 are robust even for non-machine learning techniques.

In order to confirm whether these results can be interpreted as evidence of polarization on news preferences among users, we need additional estimation:

$$|slant_{ij} - antislant_{ij}| = \alpha_j^{diff} + \gamma^{diff} customized_{ij} + \eta^{diff} X'_{ij} + \omega_{ij}^{diff}, \tag{6}$$

where an outcome variable is the absolute difference of slant indices between Hindu and anti-Hindu nationalism. If $\gamma^{diff} > 0$, polarization on news preferences become more severe in the treatment group than randomized news circulation: a user who seeks slant news on Hindu nationalism dislikes to consume news on anti-Hindu nationalism. On the other hand, if $\gamma^{diff} \leq 0$ we cannot say that such polarization is likely to move ahead. The estimated coefficient γ^{diff} is 1.293 for which the standard error is 0.384.[7] So the former inference holds in this context.

Finally, my data allows me to investigate heterogeneity impacts of AI-customized news circulation on average slants toward Hindu and anti-Hindu nationalism. I examine the heterogeneity in terms of a user's baseline belief on Hindu nationalism. In the baseline survey, I conducted a questionnaire survey in which I displayed ten "real" news articles. Each of five news articles has different tones on Hindu or anti-Hindu nationalism from

[7] A result table for this estimation is available upon request.

"strongly agree" to "strongly disagree". Then a user was required to evaluate each news article stance on a one-to-five scale: one is "strongly favor" and five is "strongly against". Using these evaluation results, I construct each user's baseline belief on Hindu and anti-Hindu nationalism. Table 3 shows heterogeneity impacts based on such a user's baseline belief. I split a full sample into three sub-samples for each slant index, i.e., six sub-samples in total: Column (1) and (4) consist of people who strongly favor Hindu and anti-Hindu nationalism, respectively. By contrast, Column (3) and (6) consist of people who are strongly against Hindu and anti-Hindu nationalism, respectively. Those who are identified as being with neutral ideology are included in Column (2) and (5). The estimation results show that AI-customized news circulation offers biased news only if a user has extreme ideology at the beginning because the coefficients for "strongly favor" are always positive and statistically significant while the coefficients for "strongly against" are negative and statistically significant. This is another evidence that AI-customized news circulation aggravates polarization on news preferences.

5.2 Impacts of Repeated Exposure to Slant News

Next, I turn to quantify the impacts of receiving highly slant news. To do so, I exclusively focus on the sub-sample for the control group, i.e., randomized news circulation, and estimate the following equations:

$$\Delta y_{ij} = \theta_j + \omega highly\ slant_{ij} + \delta X'_{ij} + \varepsilon_{ij}, \tag{7}$$

$$\Delta y_{ij} = \theta_j^{anti} + \omega^{anti} highly\ antislant_{ij} + \delta^{anti} X'_{ij} + \varepsilon_{ij}^{anti}, \tag{8}$$

where Δy_{ij} denotes the difference of experimental outcomes for user i in district k between the baseline and the endline surveys. $highly\ slant_{ij}$ and $highly\ antislant_{ij}$ denote the number of highly slant news toward Hindu and anti-Hindu nationalism that user i receives during the experimental period, respectively. If the slant index surpasses the 75th percentile, its news is defined as being highly slant.

Table 4 presents estimation results of the above equations. Outcomes of Column (1), (2), and (3) are outgroup discrimination measured by other-other allocation games for all religious and caste categories, only upper religious and caste categories, and only lower religious and caste categories, respectively. An outcome for Column (4) is spitefulness measured by a third-party punishment game while outcomes for Column (5) and (6) are political preferences toward Hindu and anti-Hindu nationalism measured by an other-other allocation game among "real" political candidates, respectively. Panel A and B

examine the impacts of news bias toward Hindu nationalism while Pane C and D examine the impacts of news bias toward anti-Hindu nationalism. Panel A and C demonstrate that the number of highly slant news itself doesn't affect the degree of discrimination at all in both cases (Hindu nationalism and anti-Hindu nationalism). However, the result prominently changes if I decompose a key variable of the number of highly slant news into two variables: the number of highly slant news with two or three positive comments or the number of highly slant news with other variations of comments such as two neutral comments and one negative comment. Note that the latter variable includes the case without any comments. Furthermore, remember that if a news article is randomly assigned to the treatment where a user can see others' comments, it always comes to a user's phone with *three* comments. Therefore, when a user faces at least two positive comments, these positive evaluations are likely to be dominant or the majority for him or her. Panel B and D present the coefficients of these two variables. Interestingly, the number of highly slant news significantly exacerbates outgroup discrimination when news articles are displayed with at least two positive comments. This result is replicated even for outgroup discrimination for upper and lower castes shown in Column (2) and (3). The same story is applicable for Column (4), (5), and (6) as well: people are more likely to being spiteful to other social classes and to prefer politicians having either Hindu or anti-Hindu nationalism ideology. This implies that a user is affected by slant news only if he or she thinks that the other user's view of an article seems to be the majority or representative.

Then, natural questions arise: do people shift their stance in either of two directions regardless of their original ideology if they happen to get slant news with seemingly dominant opinions? or does the baseline ideology on Hindu nationalism matter such that only people with extreme ideology are further shifted by slant news? These questions are important to give inference on the consequences of distorted news flows on ideological polarization. From here, I focus on a single outcome variable which is average outgroup discrimination measured by other-other allocation games. Table 5 shows that those who strongly favor Hindu or anti-Hindu nationalism are influenced by repeated exposure to slant news with others' positive comments because it is similar to own ideology and endorsed by others. As expected, neutral people are unaffected by them. Interestingly, if a user is strongly against anti-Hindu nationalism he or she also responds to slant news on anti-Hindu ideology by more aggressively discriminating others. These are is clean evidence that repeated exposure to slant news expands ideological segregation by pushing people with extreme ideology further to the margin.

Finally, how can we avoid such polarization? I also capture information on whether people are aware of their own position in terms of Hindu and anti-Hindu nationalism at the baseline and the endline surveys on a one-to-five scale. I define that a user is sophisticated if he or she correctly understands an own scale of two ideologies before the intervention. Then I utilize this baseline recognition of own ideology, to examine the heterogeneity impacts of slant news on discrimination. Table 6 illustrates stark contrast between sophisticated and non-sophisticated users: when it comes to the impacts of highly slant news with at least two positive comments, sophisticated users are totally unaffected by it while non-sophisticated users get to have more discriminatory preferences. This implies that the educational intervention which makes people correctly realize on their own ideology might contribute to avoid ideological segregation.

5.3 Identifying Mechanisms

Here, I specify the underlying mechanism behind why people's discriminatory and political preferences are affected by slant news only if it comes with positive comments posted by others. There are two candidates to explain that phenomenon: social image motivation and confirmation bias. A social-image or social reputation hypothesis explains that people sensitively respond to slant news with other users' comments because they are afraid that their own social image might be deteriorated if their ideology is distant from a general opinion in a society. So far, I find little supporting evidence on this hypothesis. According to Table 5, those who are already in favor of extreme Hindu or anti-Hindu nationalism are encouraged by slant news with positive comments, while those who are strongly against at first don't experience any change on their discrimination preference. If a social-image hypothesis holds, "strongly-against" people should more keenly adjust their ideology to opposite nationalism since it is quite far from the information that they receive. In other words, "strongly-against" people are supposed to be subject to stronger peer pressure than others.

Before rejecting the social-image hypothesis, I also check the heterogeneity impacts varied by a scale of online social networks. Using my mobile application, I can extract each user's social network information: how many friends are in online SNS services? I find no statistically significant results on heterogeneity impacts by a scale of online social networks: the number of highly slant news with positive comments for an isolated user and a user connected with many people have indistinguishable coefficients which are not

statistically different.[8] So a social-image or social reputation hypothesis doesn't explain the underlying mechanism in my context.

Next, I explicitly examine the confirmation bias hypothesis: users accept social media news and others' opinions as being representative only if the degree of their slant is similar to or on the same side of own belief. So far, I have found consistent evidence on this hypothesis: Table 5 illustrates that people only respond to highly slant news when their ideology is similar to bias in news articles. Here, I present additional evidence by using a regression discontinuity design. First, I estimate an individual slant index which reflects favoritism toward Hindu vs. anti-Hindu nationalism, utilizing user's comments during the last week right before the endline survey. More specifically, I calculate this individual slant index as the consistency or the distance between user's comments and Hindu or anti-Hindu nationalism by applying the Jensen-Shannon distance based on LDA methods as explained above. Second, in the endline survey, the app displays twenty real news articles on users' mobile phones and asks users to answer which of news articles are actually circulated to them. A half of them are news articles which are categorized as being slant to Hindu nationalism while the rest of them are categorized as being slant to anti-Hindu nationalism. Moreover, within each of ten news articles, a half of them are articles which were actually circulated to a user while another half of them are circulated to other users. Finally, the difference between the number of correct answers on Hindu-nationalistic news articles and the number of correct answers on anti-Hindu-nationalistic news articles is used as an outcome variable on a sharp RDD analysis. Figure 2 shows a graphical result of this investigation. Obviously, at the threshold where an individual slant index is equal to zero meaning a neutral stance toward nationalism, there exists a discontinuous increase of an outcome. This implies that people having non-neutral ideology may remember news articles selectively to reinforce their baseline expectations or belief. Again, this result corroborates a confirmation bias hypothesis as a underlying mechanism to explain the impact of slant news.

6 Conclusion

This paper is aimed at investigating whether social media which is characterized by AI-customized news circulation facilitates users to receive more biased news. To test this, I developed the original mobile application and released it to potential users in South India. Econometric analyses demonstrate that AI-customized news circulation actually increases

[8] An estimation table is available upon request.

the probability to receive distorted news which is similar to own ideological disposition. Furthermore, I reveal that if a user is more exposed to biased news he or she is more likely to have outgroup discrimination. This implies that AI-customized news circulation induces ideological polarization in terms of Hindu nationalism. It is worth noting that this negative impacts of social media are realized only if a user can confirm other users' positive comments on slant news. This means that we might be able to avoid such polarization by showing diversified opinions on each news. Furthermore, I also find that people with sophisticated recognition on own ideology are unaffected by such slant news. This suggests that some interventions making a user understands his or her ideological stance correctly might be able to prevent ideological polarization. Finally, the additional examination identifies that confirmation bias rather than social image motivation drives the impacts of slant news on discrimination and political preferences.

References

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., & Zhuravskaya, E. (2015). Radio and the rise of the nazis in prewar germany. *The Quarterly Journal of Economics*, *130*(4), 1885–1939.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, *107*(3), 797–817.

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, *102*(5), 841–877.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In *Handbook of economic field experiments* (Vol. 1, pp. 309–393). Elsevier.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–57.

DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., & Zhuravskaya, E. (2014). Cross-border media and nationalism: Evidence from serbian radio in croatia. *American Economic Journal: Applied Economics*, *6*(3), 103–32.

Enikolopov, R., Makarin, A., & Petrova, M. (2017). Social media and protest participation: Evidence from russia. *Available at SSRN 2696236*.

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

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63–87.

Gentzkow, M. (2006). Television and voter turnout. *The Quarterly Journal of Economics*, *121*(3), 931–972.

Gerber, A. S., Karlan, D., & Bergan, D. (2009). Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, *1*(2), 35–52.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Sunstein, C. R. (2018). *# republic: Divided democracy in the age of social media*. Princeton University Press.

Table 1: Impacts of Customized News Circulation on Variables Estimated by Machine Learning Methods

Machine Learning Methods (LDA)	(1)	(2)	(3)	(4)
Panel A: Slant toward Hindu nationalism	Average	Standard deviation	Max	Min
Customized news circulation (CNC)	0.0128	-1.175***	1.249***	-1.388***
	(0.0155)	(0.211)	(0.238)	(0.277)
Other controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Obs	25,774	25,774	25,774	25,774
Machine Learning Methods (LDA)	(5)	(6)	(7)	(8)
Panel B: Slant toward anti-Hindu nationalism	Average	Standard deviation	Max	Min
Customized news circulation (CNC)	0.0183	-1.012***	1.441***	-1.291***
	(0.0199)	(0.0213)	(0.241)	(0.201)
Other controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Obs	25,774	25,774	25,774	25,774

Note: ***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Table 2: Impacts of Customized Circulation on Variables Estimated by non-Machine Learning Methods

Non-Machine Learning Methods	(1)	(2)	(3)	(4)
Panel A: Slant toward Hindu nationalism	Average	Standard deviation	Max	Min
Customized news circulation (CNC)	0.00921	-0.284***	0.103***	-0.114***
	(0.0138)	(0.0258)	(0.0389)	(0.0522)
Other controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Obs	25,774	25,774	25,774	25,774
Non-Machine Learning Methods	(5)	(6)	(7)	(8)
Panel B: Slant toward anti-Hindu nationalism	Average	Standard deviation	Max	Min
Customized news circulation (CNC)	0.00843	-0.212***	0.100***	-0.119***
	(0.0152)	(0.0274)	(0.0317)	(0.0501)
Other controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Obs	25,774	25,774	25,774	25,774

Note: ***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Table 3: Heterogeneity Impacts of Customized News Circulation

Dep var: Average Slant Index	(1)	(2)	(3)
	Baseline Belief on Hindu Nationalism		
Panel A: Hindu nationalism	Strongly Favor	Neutral	Strongly Against
Customized news circulation (CNC)	1.230*	0.0294	-0.0403**
	(0.613)	(0.0258)	(0.0200)
Other controls	Yes	Yes	Yes
District FEs	Yes	Yes	Yes
Obs	4,031	17,791	3,952
Dep var: Average Slant Index	(4)	(5)	(6)
	Baseline Belief on anti-Hindu Nationalism		
Panel B: anti-Hindu nationalism	Strongly Favor	Neutral	Strongly Against
Customized news circulation (CNC)	1.081**	0.0212	-0.0501**
	(0.489)	(0.0238)	(0.0217)
Other controls	Yes	Yes	Yes
District FEs	Yes	Yes	Yes
Obs	4,532	17,182	4,060

Note: ***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Table 4: The Impacts of Highly Slant News on Discrimination

Sample: only for randomized news circulation	(1)	(2)	(3)	(4)	(5)	(6)
	Discrimination			Spiteful	Political Pref.	
	All	Upper	Lower		Hindu	anti-Hindu
Panel A: A slant for Hindu nationalism						
No. of highly slant news	0.0362	0.0571	0.0279	0.0190	0.481	0.381
	(0.0721)	(0.109)	(0.113)	(0.0245)	(0.294)	(0.388)
Panel B: A slant for Hindu nationalism						
No. of highly slant news (positive comments < 2)	0.0183	0.0285	0.0149	0.0101	0.214	0.281
	(0.0742)	(0.188)	(0.125)	(0.0138)	(0.298)	(0.213)
No. of highly slant news (positive comments ≥ 2)	0.181***	0.117***	0.104***	0.135*	0.542**	0.333
	(0.0430)	(0.0326)	(0.0358)	(0.0651)	(0.251)	(0.246)
Panel C: A slant for anti-Hindu nationalism						
No. of highly slant news	0.0280	0.0417	0.0382	0.0135	0.449	0.337
	(0.0681)	(0.132)	(0.233)	(0.0348)	(0.312)	(0.288)
Panel D: A slant for anti-Hindu nationalism						
No. of highly slant news (positive comments < 2)	0.0112	0.0133	0.0109	0.0102	0.291	0.227
	(0.0740)	(0.0188)	(0.0225)	(0.0202)	(0.296)	(0.225)
No. of highly slant news (positive comments ≥ 2)	0.147***	0.171***	0.101***	0.118*	0.522	0.484*
	(0.0235)	(0.0282)	(0.0399)	(0.0521)	(0.310)	(0.256)
Other controls	YES	YES	YES	YES	YES	YES
Obs	17,182	17,182	17,182	17,182	17,182	17,182

***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Table 5: Heterogeneity Impacts of Highly Slant News on Discrimination

Sample: only for randomized news circulation	(1)	(2)	(3)
	Baseline Belief on Hindu Nationalism		
Dep var: outgroup discrimination	Strongly Favor	Neutral	Strongly Against
Panel A: A slant for Hindu nationalism			
No. of highly slant news	0.0318	0.0371	0.0371
	(0.0483)	(0.0445)	(0.0434)
Panel B: A slant for Hindu nationalism			
No. of highly slant news (positive comments < 2)	0.0201	0.0229	0.0219
	(0.0635)	(0.0483)	(0.0151)
No. of highly slant news (positive comments ≥ 2)	0.131***	0.0117	0.0309
	(0.0410)	(0.0396)	(0.0182)
Other controls	Yes	Yes	Yes
District FEs	Yes	Yes	Yes
Obs	6,125	11,872	5,747
	Baseline Belief on anti-Hindu Nationalism		
Dep var: outgroup discrimination	Strongly Favor	Neutral	Strongly Against
Panel C: A slant for anti-Hindu nationalism			
No. of highly slant news	0.0322	0.0337	0.0351
	(0.0482)	(0.0474)	(0.0433)
Panel D: A slant for anti-Hindu nationalism			
No. of highly slant news (positive comments < 2)	0.0298	0.0213	0.0201
	(0.0610)	(0.0412)	(0.0452)
No. of highly slant news (positive comments ≥ 2)	0.137***	0.0177	0.0299*
	(0.0435)	(0.0381)	(0.0149)
Other controls	Yes	Yes	Yes
District FEs	Yes	Yes	Yes
Obs	5,025	11,454	6,429

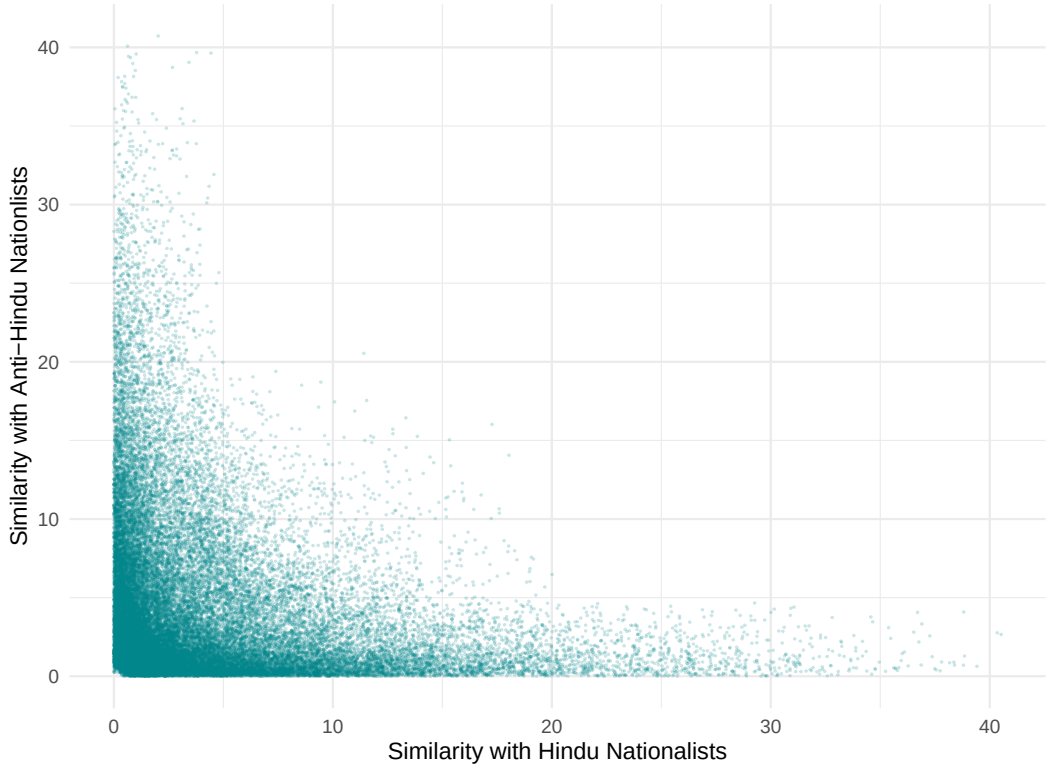
Note: ***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Table 6: Heterogeneity Impacts between Sophisticated and Non-sophisticated Users

Sample: only for randomized news circulation	(1)	(2)
	Recognition of Own Ideology	
Dep var: outgroup discrimination	Sophisticated	Non-sophisticated
Panel A: A slant for Hindu nationalism		
No. of highly slant news	0.0304	0.0310
	(0.0428)	(0.0438)
Panel B: A slant for Hindu nationalism		
No. of highly slant news (positive comments < 2)	0.0197	0.0109
	(0.0538)	(0.0429)
No. of highly slant news (positive comments ≥ 2)	0.0298	0.128***
	(0.0485)	(0.0325)
Other controls	Yes	Yes
District FEs	Yes	Yes
Obs	6,281	19,493
Panel C: A slant for anti-Hindu nationalism		
No. of highly slant news	0.0322	0.0319
	(0.0447)	(0.434)
Panel D: A slant for anti-Hindu nationalism		
No. of highly slant news (positive comments < 2)	0.0210	0.0111
	(0.0531)	(0.412)
No. of highly slant news (positive comments ≥ 2)	0.0232	0.172***
	(0.0473)	(0.0344)
Other controls	Yes	Yes
District FEs	Yes	Yes
Obs	6,281	19,493

Note: ***, **, and * denote statistical significance in 1%, 5%, and 10%, respectively.

Figure 1: Plots for Estimated Slant Indices



Note: Each dot represents a news article that was circulated to at least one user during the experimental period. Similarity with Hindu nationalists denotes the Jensen-Shannon distance between each news article and the corpus of Hindu nationalists while similarity with anti-Hindu nationalists denotes the Jensen-Shannon distance between each news article and the corpus of anti-Hindu nationalists.

Figure 2: A Biased Memory as Evidence of Confirmation Bias