Local Elections, Leader Identity and Hate Speech in Rural India

Aarushi Kalra*

August, 2022

Abstract

How does the identity of local political leaders change public and private expression of political opinions? Using data from a very popular Indian content generation platform, we analyze if social media users are more likely to engage with hateful content when the local leader belongs to a vulnerable minority group. We consider engagement activity of 19 million users with one million Hindi posts around the village council elections of April, 2021 in Uttar Pradesh, India. This is done to test the existence of a backlash effect when a member of the minority group gains political power. We employ a regression discontinuity design to compare anti-minority hate speech in villages where a Muslim candidate won an election with villages where Muslim candidates lost by a small margin. We find no change in engagement with toxic content in villages where the elected leader was Muslim after the election result was announced. A potential explanation is that engagement with hate is driven by electoral competition and campaigning efforts, and not the announcement of results. We also find evidence that engagement on the platform is generated by national news trends, and platform algorithms that do not take into account user location while generating customized content recommendations. This can make highly localized political conditions less salient for political engagement on social media platforms. This paper provides new evidence on political factors, like local election campaigns and identity of local political leaders, that are expected to change norms of engagement with political and toxic content on social media.

1 Introduction

There has been a sharp increase in political polarization and hate speech on social media platforms around the world (Halberstam and Knight, 2016; Boxell et al., 2020; Allcott et al., 2020). The allegations of Facebook whistleblower, Frances Haugen, the violent events of January 6, 2021, on the Capitol Hill in Washington DC , and Facebook's failure to prevent incitement of violence in Myanmar have piqued academic interest in the factors that lead to the production and propagation of hateful content on various platforms (Ledwich and Zaitsev, 2019; Huszár et al., 2022). With over 500 million social media users, India provides a highly relevant context for studying these questions as social media has grave consequences for the socio-political environment in this setting.

Social media posts are known to have provoked instances of violence, in the form of mob lynching, riots, and hate crimes in India (Mukherjee, 2020). To address this problem, researchers first need to learn the factors that cause engagement with hateful content. This helps in understanding where the problem stems from, in order to recommend appropriate policy instruments for its' regulation. These could be determined by adverse economic shocks as well as socio-political environment of users, in addition to technological factors stemming from social media platforms, like content recommender systems.

In this paper, we focus our attention on political shocks that cause changes in the social identity of local political leaders. We interrogate how the identity of local political leaders change public and private expression of political opinions on social media platforms. We study the effect of changes in leader identity

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

in village Council or *Gram Panchayat* (henceforth, *Panchayat*) elections on user engagement with hateful content, on a very popular social media platform in India. We hypothesize that in villages where a leader from the Muslim minority is elected to the village council, social media users increase their engagement with hateful content after the election results are announced. This is plausible on account of backlash from various dominant caste and religious groups in this setting (Drèze, 2020; Mitra and Ray, 2014; Jaffrelot and Kalaiyarasan, 2019).

We observe social media activity of users on ShareChat, a content generation platform with 180 million active monthly users in India. Users create and share data in 14 regional languages on the platform. Like TikTok, ShareChat is an upstream content generation app that caters to the content needs of a very large portion of the 459 million Indian WhatsApp users. This is due to ShareChat's interface with WhatsApp, which enables users to share image and video content. Therefore, we observe which users share posts on WhatsApp. Further, we identify socio-economic conditions of users by linking their geo-location to data on village council elections from the Uttar Pradesh State Election Commission (Srivastava et al., 2022).

We employ a regression discontinuity design to study differences in engagement with hate speech in villages where a Muslim candidate wins an election by a very small margin. This is done under the assumption that villages where a Muslim candidate wins the election by a small margin are similar, in observable and unobservable attributes, to villages where a Muslim candidates loses the election to a non-Muslim candidate (Hahn et al., 2001). This margin is computed for the given data according to robust methods in Calonico et al. (2014).

We do not find any differential patterns in engagement with hate speech in treated villages, after the announcement of election results. That is, a transfer of political power to Muslim candidate does not lead to more or less engagement[1] with hateful content on social media. The absence of an expected backlash effect is in part explained by the fact engagement with hateful content is driven up before the announcement of results, presumably by election campaigning efforts.

We can explain this surprising result by the fact that content engagement on the platform is largely driven by national news trends. We provide evidence that exposure to hate speech is highly correlated across villages in the state of Uttar Pradesh, and is generated by Islamophobic news items in mainstream national news outlets[2]. This is also because ShareChat's content recommendation algorithms do not customize user feed according to their geographic location, thus generating correlated exposure across space. In addition, we find that local content creators do not respond strongly to local political conditions in rural Uttar Pradesh.

The remaining paper is organized as follows: Section 2 lists out all our data sources, including construction of outcome variables and treatment groups, in addition to highlighting the setting for our rich social media network data from ShareChat. In this section, we also highlight the machine learning methods employed to process these data. Section 3 presents the main empirical strategy and offers a discussion of the methods employed. We present our main results in Section 4. We provide insights on mechanisms that might generate engagement with and exposure to hateful content on the platform in Section 5. Finally, we discuss the limitations and conclusions of our research in Section 6.

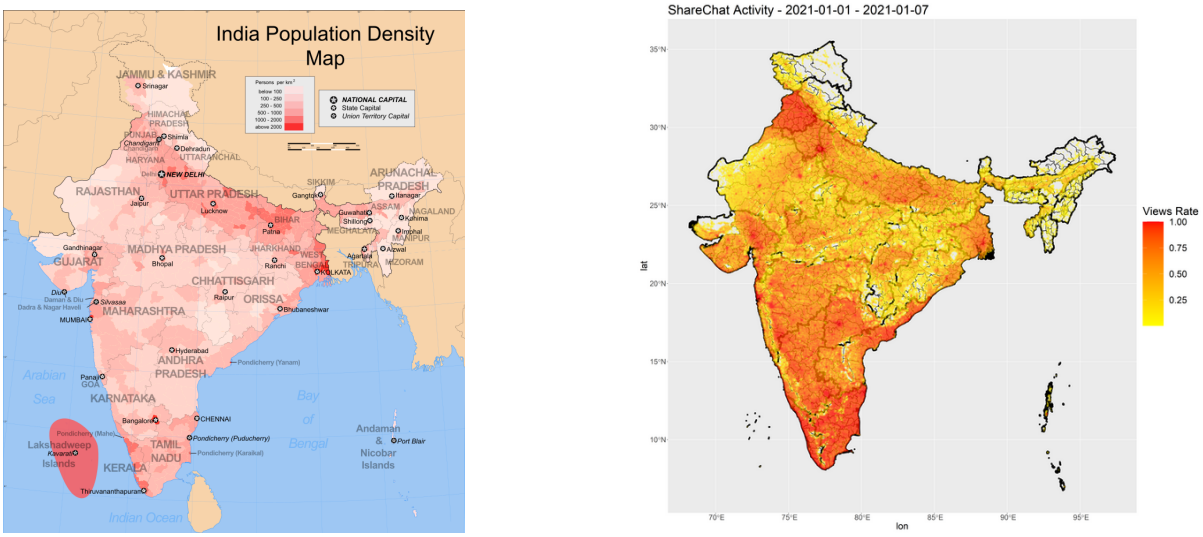
[1] We measure public engagement with content using number of *likes* in village i at time t divided by the total number of users in village v. This is because likes are observable by other users. Similarly, we measure private engagement with content using number of *shares* in village i at time t divided by the total number of users in village v. This is because the identity of a user sharing a piece of content is not observed by other users.

[2] Data on news text was collected from the twitter handle of OpIndia, a righ-wing mainstream news outlet, using the Twitter API.

2 Data

2.1 Platform: ShareChat

We use data from ShareChat, a rapidly growing social media platform in India. ShareChat is one of the most popular platforms in the country with over 180 million active monthly users, who create and share content in 14 regional languages on the platform (See Figure 1 for ShareChat's reach). The platform caters to the content needs of a very large portion of the 459 million Indian WhatsApp users[3]. Apart from publicly observable posts, we are able to view which users share these posts to WhatsApp. This is, therefore, a rich setting to understand how various social groups interact with each other, and the factors that lead to hateful content generation and sharing. We consider engagement activity of 19 million users with a million Hindi posts around the village council elections of April, 2021 in UP.



(a) Population denisty map of India

(b) Average viewing of content on ShareChat in 2021.

Figure 1: Comparing population density across Indian districts with average activity on ShareChat for one week in 2021.

Of these modes of engagement: liking, reposting and commenting features are public, meaning users on ShareChat can observe the above actions on the content. The downloading and sharing features are private, meaning that other users cannot observe who shared what to WhatsApp and other platforms. We observe all these forms of interaction on ShareChat, both public and private. Crucially, we rely on 'viewing' data at the user-post level to measure a user's exposure to various forms of content on the platform.

ShareChat is comparable to TikTok in its content generation and engagement features and attracts a large portion of young voters among the urban and rural poor in India. This makes such analysis especially important as little is known about political behavior of this demographic in India.

[3] https://techcrunch.com/2021/01/11/youtube-and-whatsapp-inch-closer-to-half-a-billion-users-in-india/



Figure 2: User view of home screen, for a Hindi user, under the topic of 'Political Discussions' The first tab (from the left) on the screen is called 'Trending'.

Examples of popular political content in Hindi (along with corresponding English transliterations) are provided in the word cloud in Figure A.1. It is worth noting that highest weight (as computed by the highest number of appearances in political posts on the platform) is given to the text *Jai Shri Ram* (or, Hail Lord Ram)– which is a deliberately polarizing slogan of the Hindu Nationalist organizations in India. Anti-minority hate speech forms the main outcome variable of the research questions we seek to answer. Since hate speech classification is a key part of our research, we use cutting edge tools in Machine Learning for this task. These methods are described below.

2.2 Outcome: Hate Speech/ Toxic Speech

We analyze the effect of change in identity of local leader on engagement with hateful content. We construct the outcome variable of interest by classifying text on ShareChat posts that were created around the Panchayat elections in UP. We follow multiple approaches to arrive at accurate classification of content into two categories: Hate Speech or non-Hostile. Firstly, we employ the top performing Machine Learning Algorithm[4] at the conference for Hate Speech and Offensive Content Identification in Indo-European Languages, HASOC[5]. We plot trends in engagement with Hate Speech on ShareChat using this measure in Figure 3. Since this algorithm is trained on a set of Hindi *tweets* (and not ShareChat posts), the classification exercise renders a high degree of false positives. Moreover, classification of hate speech is inherently tricky and is highly prone to bias due to the training dataset's composition. Ideally, political and social contexts should inform what is deemed as hateful.

[4]We found the accuracy to be 65% on political ShareChat posts. See Das et al. (2022) for a detailed analysis of various Hindi hate speech classifiers.

[5]HASOC2020, (Mandl et al., 2020)

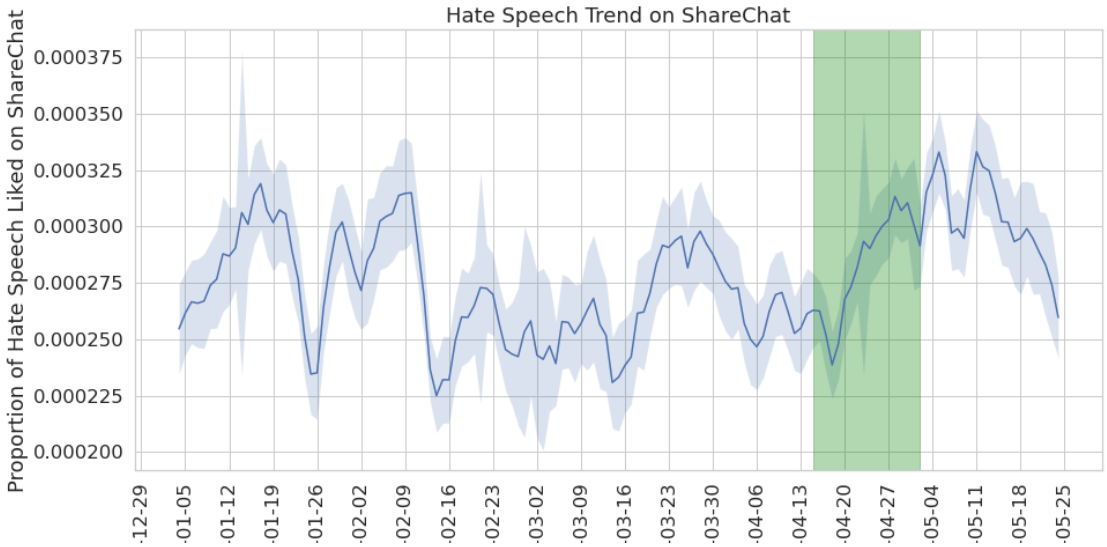


Figure 3: Trends in Engagement with Hate Speech in Uttar Pradesh, India. Shaded region depicts timeline for Panchayat elections in UP, conducted in four phases in April, 2021.

Secondly, we employ a different outcome measure (engagement with posts close to topics containing the slogan *Jai Shri Ram*, or Hail Lord Ram). We use a Latent Dirichlet Allocation (LDA) Model to identify primary topics of discussion in the posts classified as hateful by the HASOC algorithm. We perform our analysis using posts closest to *Jai Shri Ram* as a concrete measure of Hate Speech, as this is the rallying cry of Hindu Nationalist forces in India[6]. This measure tracks very closely with the measure produced using the HASOC classifier, and therefore, suffers from the same limitations.

In recent years, automated detection of hate speech content has received significant attention from the Natural Language Processing (NLP) community. While hate speech classification remains a challenging task for Hindi and multilingual text, there are multiple promising off-the-shelf models which out-perform the two approaches mentioned above. Perspective API[7] is one such off-the-shelf model that provides continuous labels for various attributes such as 'toxicity', 'severe toxicity', 'insult', 'identity attack' and 'profanity'. Out of all attributes offered, only toxicity is available for Hindi (Devanagari script), while others can be utilized with English translated versions of the text. Given the slew of different types of hate speech detection models being available, there have been efforts to evaluate them that have revealed significant weaknesses in all models, especially for multilingual text classification (Das et al., 2022). These efforts also highlight the successes of Perspective API and BERT-based classifiers for these tasks.

Finally, after experimenting with various classifiers mentioned above, we found the most success with Perspective API. While the outcome variable for Perspective API isn't 'Hate', the continuous 'toxicity' measure can be binarized to achieve better performance on the hate speech classification task than other models, especially for Hindi (See Figure 4 below). Moreover, the variety of continuous measures provided by Perspective API give flexibility and power in studying different features of hate speech on social media platforms.

[6]https://foreignpolicy.com/2020/02/13/jai-shri-ram-india-hindi/
[7]`https://perspectiveapi.com/`

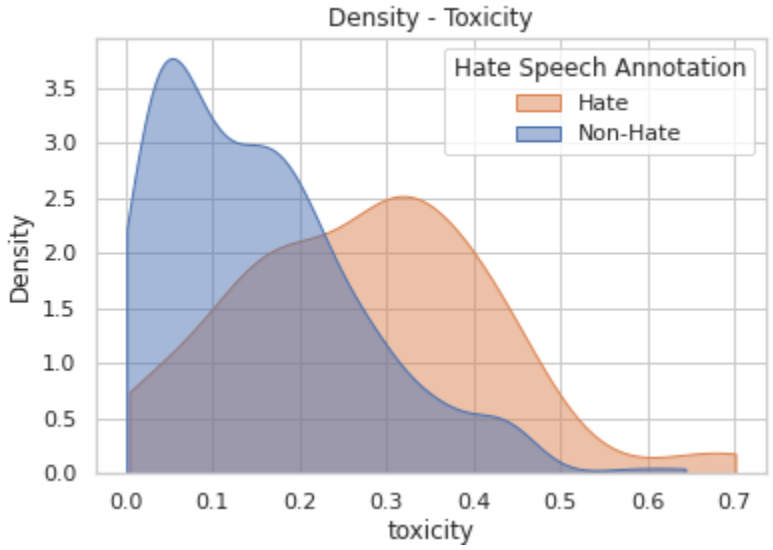


Figure 4: Distribution of toxicity scores, grouped by manual annotation for a post being hateful or not.

Figure 4 shows the distribution of toxicity score (a continuous measure lying between zero and 1) for posts that are manually annotated as hateful or not. The manual annotation was conducted by a team of researchers at Brown University, following definitions for Hindi-English code-mixed social media text in Santosh and Aravind (2019). We further test the performance of this method by comparing discrepancies between hate speech prediction from the classifier (by assigning value 1 to a post if the toxicity score is above a certain threshold), against our annotation. We are able to increase the accuracy of the algorithm by fine-tuning the threshold in Table B.1. We provide example posts that generate false positives in Table B.2.

2.3 Treatment: Panchayat Elections

Panchayat elections are local council elections in rural India. These elections consist of 500 to 10,000 adults voting to elect the representatives in around 60,000 villages[8] in the state of Uttar Pradesh alone (MoSPI, 2011). While election candidates' data is openly available on Uttar Pradesh State Election Commission website, it is not easily accessible. Moreover, we were unable to find accurate public geolocation data for villages in the state. This motivated us to scrape the state website to obtain the data, and develop a method to geolocate the villages using the Google Maps API to allow us to match them with our social media data consisting of the user's activity and their geo-location. 2021 election data was then joined with 2015 election data using both the names of the Gram Panchayat and Zila (district). Only villages that held panchayat elections in 2015 and 2021 are included in the dataset. Our final dataset consists 56,462 geolocated villages across Uttar Pradesh.

We also classified the religion of each candidate as Muslim or Non-Muslim using a neural network model we trained on a manually annotated subset of our data containing names in Hindi classified as Muslim or Non-Muslim[9]. Table 1 summarizes trends in Panchayat election results for Muslim candidates based on our classification, in villages that we matched to the ShareChat database.

[8]We use the term Panchayat and village interchangeably, although the administrative unit of Gram Panchayat may consist of a cluster of more than one village.

[9]We found the accuracy of the model to be 95%

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

Table 1: Muslim candidates (winners and runners) as percentage share of all candidates in 2021 and 2015 *Panchayat* elections. The reported averages are over all villages matched to the ShareChat database. Standard deviations are reported in parenthesis.

Based on the religious identity of the incumbent and newly elected leader, we partition our set of villages into four groups depicted in Table 2 where the first element of every tuple depicts the identity of the incumbent (winner in 2015 elections), and the second element does the same for the winner of the 2021 election.

	Winner	
Incumbent	(Muslim, Muslim)	(Muslim, Non-Muslim)
	(Non-Muslim, Muslim)	(Non-Muslim, Non-Muslim)

Table 2: Partitions for villages, based on the religious identity of incumbent and winner, to analyze trends in toxic speech with respect to date of polling in 2021 *Panchayat* elections as well as date of announcement of election results.

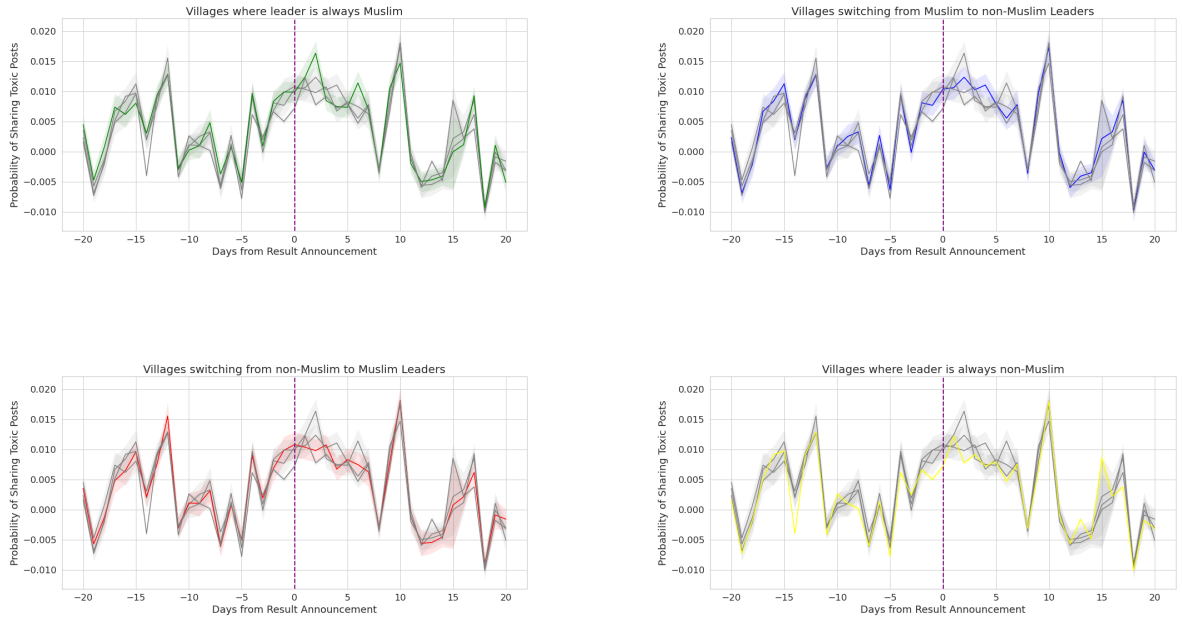


Figure 5: Trends in Sharing toxic content, relative to date of announcement of UP *Panchayat* elections in 2021. Each of the subfigures highlight trends in engagement with toxic speech for the four types of villages. Village type is determined by religious identity of winner and incumbent in Table 2. Average engagement with toxic speech at the village level is residualized with respect to day of the week.

If there are significant backlash effects, we would expect villages with a Muslim winner in 2021 to witness higher levels of hate speech. In particular, the resentment is expected to be higher in places where political power switches from a non-Muslim leader to a Muslim leader. To test this, we plot average toxicity scores on shared posts across the four types of villages around the date of announcement of winners in Figure 5. We also check for pretrends in the analysis.

Figure 5 shows an increase in engagement with toxic speech around ten days after the announcement of results. However, no one group of villages displays trends that are discernible by the religious identity of the winner or the incumbent. Further, we check if social media activity differs in the four types of villages with respect to date of election. This is because, the campaigning period is more politically charged and may provide greater salience to religious identity of the incumbent, as well as that of the candidates. Therefore, we also look for an increase in engagement with toxic posts during the campaigning period, prior to the elections.

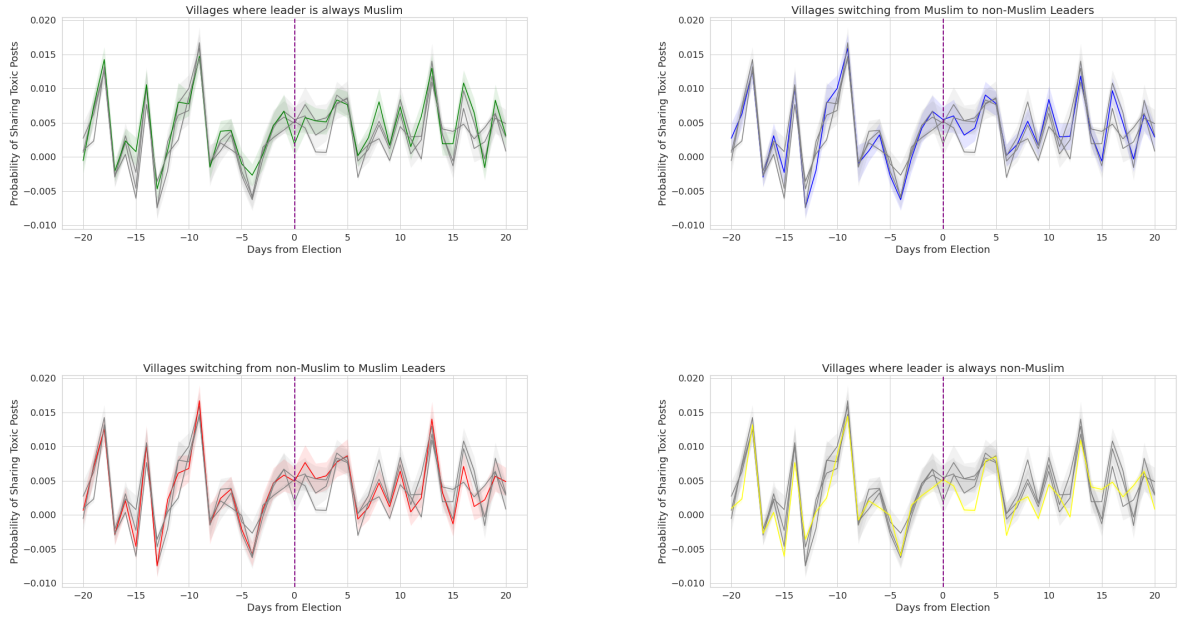


Figure 6: Trends in Sharing toxic content, relative to date of polling for UP *Panchayat* elections in 2021. Each of the subfigures highlight trends in engagement with toxic speech for the four types of villages. Village type is determined by religious identity of winner and incumbent in Table 2. Average engagement with toxic speech at the village level is residualized with respect to day of the week.

Polling took place in four phases across the districts of Uttar Pradesh in April, 2021, this provides variation in data of elections in the state[10]. However, Figure 6 shows that even though there is increased engagement with toxic content prior to polling, there is a high degree of correlation between trends in engagement with toxic speech across the four types of villages.

[10]This variation is less than desirable because the polling dates in the four phases of election were very close to each other.

3 Empirical Strategy

We employ regression discontinuity design (Calonico et al., 2014) to analyze user engagement with toxic content on social media when a Muslim political candidate wins the Panchayat election race against a non-Muslim candidate by a small margin. The control group is given by the set of villages where a Muslim candidate just loses the race to a non-Muslim candidate. We interrogate if the election of Muslim leaders leads to backlash against minority communities (Mitra and Ray, 2014; Jaffrelot and Kalaiyarasan, 2019; Drèze, 2020) in the form of toxic speech on social media. More concretely, we wish to estimate the following local linear regression:

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

where, y_{vt} is the average toxicity score of shared posts, in village v, on date t. We evaluate the outcome variable in the pre or post election announcement period separately. In particular, the outcome variable is constructed as the average of the toxicity score on shared posts for every politically active user in village v, in the relevant time period[11]. More concretely, the outcome variable, probability of sharing toxic posts, is constructed as a continuous variable as shown below:

$$y_{vt} = \sum_{i \in \mathcal{I}_v} \left(\frac{\sum_{j \in (J_{vt} \cap H)} y_{ijvt}}{|J_{vt}|} \right) \Big/ |\mathcal{I}_v| \tag{2}$$

The set of all politically active users in village v is denoted by $\mathcal{I}_v$, and the set of all posts that users in village v engage with at time t is given by J_{vt}. We average over the toxicity scores for each of these posts, $j \in (J_{vt} \cap H)$, as well as the fixed number of politically active users in village v, to arrive at the desired measure of our outcome variable.

Further, in (1), VM_v^{muslim} measures the vote margin of Muslim candidates $\in [-l, h]$ which denotes the optimal bandwidth of vote margin with respect to non-Muslim candidates. Note that we only consider elections where the identity of the candidate is salient. This means that our sample of elections only consists of villages where the winner or the runner-up was a Muslim. As a result, the measure of vote margin for Muslim candidates is necessarily in comparison with non-Muslim candidates. We choose optimal bandwidths following robust procedures in Calonico et al. (2014). The function $f(M_v)^{\text{muslim}}$ allows us to flexibly control for Muslim vote margins across elections in sample villages. Finally, X_{vt} denotes average user attributes in village v at time t, and δ_t denotes date fixed effects.

The validity of this research design relies on the assumption that potential outcome functions are continuous at the cutoff (Hahn et al., 2001).

$$\text{E}[y_{vt}(d) | VM_v^{\text{muslim}} = z] \text{ is continuous at 0 for } d = 0, 1 \tag{A1}$$

Recall that,

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

With this assumption, we now estimate the effect of the religious identity of elected leaders in village councils in Uttar Pradesh, where a Muslim candidate wins an election agains a non-Muslim candidate by a small margin. The main outcome of interest is the average toxicity score on shared posts in a given village on a particular day.

[11] We define a political active user in the election period as users who were exposed to political content a month before the Panchayat elections.

4 Results

We bring our methods to the data to test the existence of backlash effect in the form of anti-minority toxic speech on social media. We employ a regression discontinuity design in (1) to compare toxic speech in villages where a Muslim candidate won the election with villages where a Muslim candidate lost by a small margin, after the election results were announced. In this way, we compute the local average treatment effect of minority winners on patterns of toxic speech, by comparing with control villages where minority candidates attain the second position. This is done by assuming that control villages, where margin of victory for Muslim candidate is within an optimal bandwidth (computed as in Calonico et al. (2014)), have similar unobservable attributes as the villages that were treated, **??** and A1. We also require that there are no pre-trends in the data, that is the outcome variable does not differ in treated villages prior to the announcement of results.

We normalize the outcome variable, toxicity score on shared posts, by dividing it with y_{v0} which is average toxicity score in village v during some baseline period. The chosen baseline period in this analysis ends one month before the election, and includes posts from February and March, 2021. Therefore, we measure the percentage *change* in engagement with toxic content, in treatment and control villages, from some baseline engagement with such content. Furthermore, we analyze toxic engagement in sample villages 20 days before and 20 days after the announcement of election results separately.

We find a slight increase in toxicity scores of shared posts in treated villages in Figure 7 before and after the announcement of election results. However, it is evident from these figures that we cannot be confident about a discontinuous increase in treated villages, once the election results are announced.

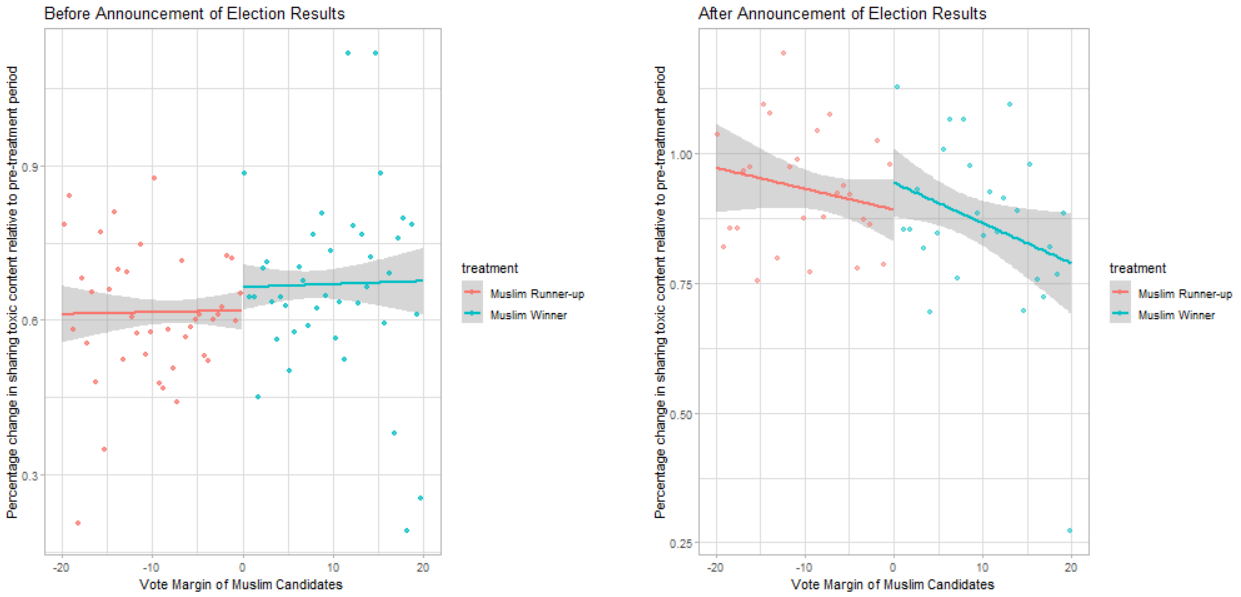


Figure 7: Percentage change in toxicity of shared posts in treated (Muslim winner) and control (Muslim runner-up) villages. Left panel shows results from RDD in the pre-announcement period. Right panel shows the same for post-announcement period.

Due to computational constraints, we estimate the magnitude of this discontinuity only for a random sample of all villages where elections were considered to be salient. Table 3 shows relevant magnitudes when this specification was run on 10% of the sample villages. Column (1) shows that there was an 11% increase in toxicity of shared posts from the baseline during the pre-announcement period. However, this increase is not statistically significant.

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

Table 3: Local linear regression results on 10% random sample, pre and post result announcement, on subsample of elections salient by religious identity. Salience requires that a Muslim candidate should finish first or second in the electoral race. Outcome variable is log difference between average toxicity score of shares in a village and average toxicity score on shares during baseline period (Febraury-March, 2021). Standard errors in paranthesis. Bias corrected and Robust 95% confidence intervals reported.

In the post-announcement period, there toxicity of shared posts increases by 17%, and this increase is also not statistically significant. Therefore, we cannot reject the hypothesis of a null backlash effect from electing Muslim leaders. We do not find evidence of increased engagement with toxic content in villages where the elections resulted in a transfer of political power to a member of the minority community.

5 Discussion

In order to understand our results, we propose some mechanisms that drive the conditions for production and propagation of toxic content on the platform.

5.1 Correlated Patterns of Exposure

We find that, irrespective of local conditions and identities of incumbents or winners, users are exposed to similar content on the platform. In Figure 8, we unpack the contents of posts that are classified as toxic, in order to understand to nature of content that users are being exposed to. Here, we model topics on our text data, and partition the sample of posts into two: posts that have high degree of proximity to the topic containing the polarizing phrase *Jai Shri Ram*, and posts that do not belong to this topic. We find that the patterns of exposure, by way of views on posts, to this polarizing phrase are practically indistinguishable in the four types of villages.

Engagement with hateful content is conditional on having viewed it in the first place. This may be the reason why local political conditions do not matter for engagement, as users view similar content in the entire state of Uttar Pradesh. Since viewing alters the sample space of posts a user can engage with in both treated and control villages, our effects (or the lack, thereof) may be confounded by exposure to content on the platform.

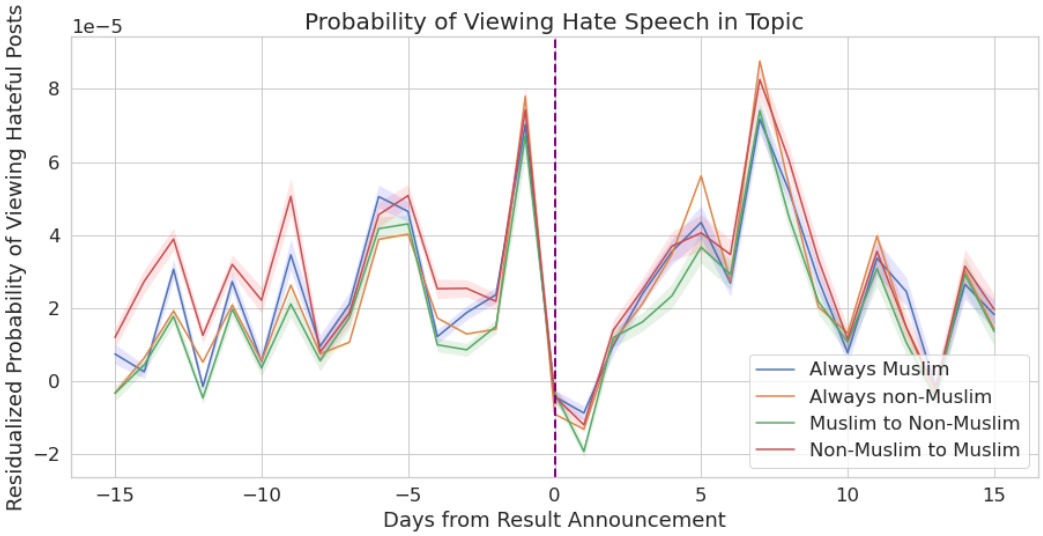


Figure 8: Trends in exposure to hateful content, relative to date of announcement of UP *Panchayat* elections in 2021. Each of the subfigures highlight trends in engagement with toxic speech for the four types of villages. Village type is determined by religious identity of winner and incumbent in Table 2. Posts are classified as hateful in accordance with their proximity to the phrase *Jai Shri Ram* (or, Hail Lord Ram). Average engagement with hate speech at the village level is residualized with respect to day of the week.

Note that exposure to content is driven by platform recommender systems or algorithms that are customized according to user preferences, and previous engagement with content on the platform. Note, however, that these algorithms do not take into account user location while generating content recommendations. This explains why patterns of exposure are so correlated across villages in Uttar Pradesh, and are possibly immune to local political conditions.

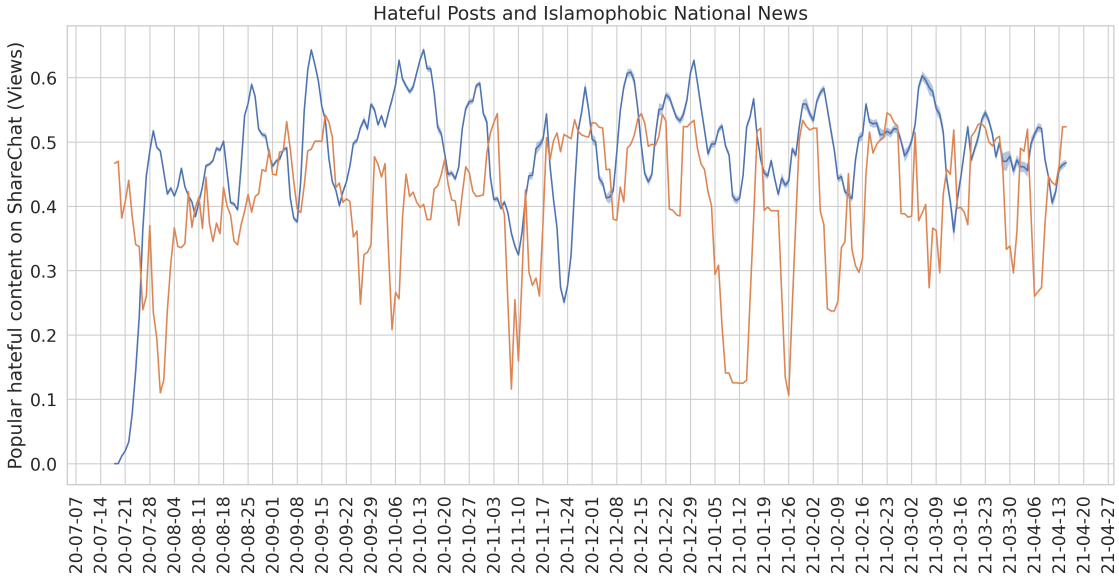


Figure 9: Trends in exposure to hateful content, by way of views, on ShareChat (blue) compared with engagement with OpIndia posts on Twitter (orange). The two trends show a high degree of correlation.

5.2 Production of and Exposure to Hate

Given the high degree of correlation between content that users are exposed to across villages, we ask: what are the chief drivers of engagement with hate speech? In order to check if these factors are determined locally or nationally, we studied engagement with posts by @OpIndia_com on Twitter. OpIndia[12] is a right-wing national news portal known for publishing and spreading anti-Muslim commentary, and spreading disinformation. Data about OpIndia posts on Twitter was collected using the Twitter API and engagement was studied using likes on Twitter. Figure 9 shows engagement with hate speech on ShareChat (using the HASOC classifier) and engagement with posts by the Twitter handle @OpIndia_com. We see a high degree of correlation between trends in engagement with hateful content on the platform and engagement with Islamophobic content in national news on Twitter. This suggests that engagement with hate speech is primarily driven by national, rather than local news and events.

Local political conditions could drive hateful engagement if they were affecting content *creation* locally. Figure 10 demonstrates that we do not see any differential patterns in production of hateful content in rural Uttar Pradesh by village type. Trends in content creation are highly correlated across different types of villages. This further provides evidence that national news is a more important factor in terms of driving engagement with toxic content, than local political conditions.

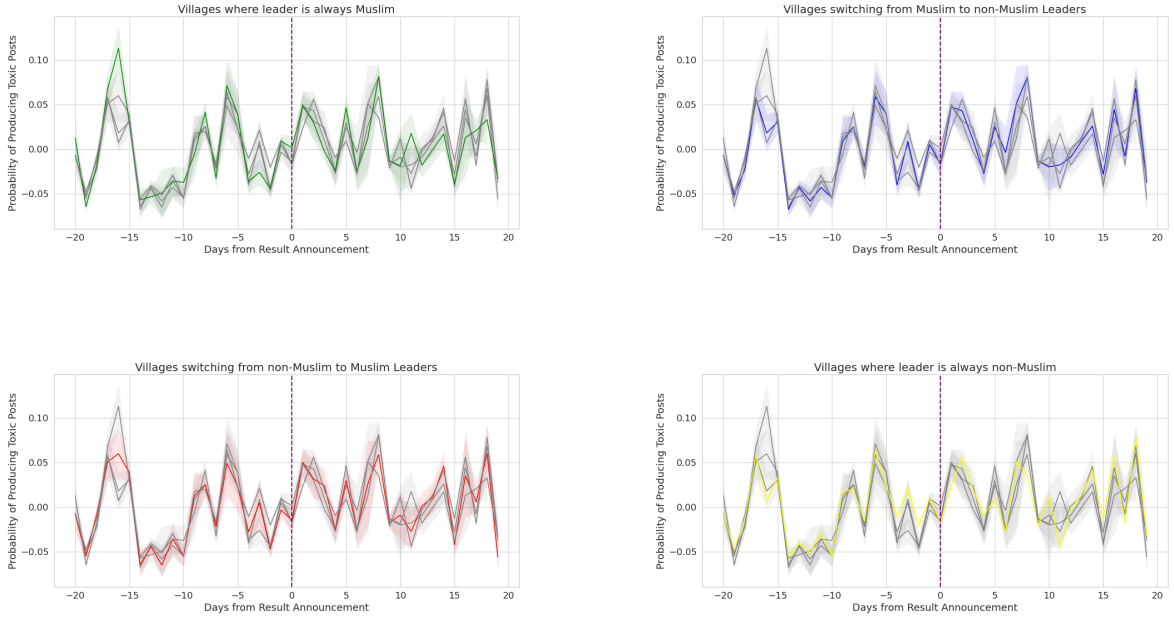


Figure 10: Trends in production of toxic content, relative to date of result announcement for UP *Panchayat* elections in 2021. Each of the subfigures highlight trends in engagement with toxic speech for the four types of villages. Village type is determined by religious identity of winner and incumbent in Table 2. Average engagement with toxic speech at the village level is residualized with respect to day of the week.

5.3 Election Cycle

It is plausible that that elections and concomitant campaigning activity may be driving hateful engagement. This is because Figure 7 implies the presence of some pre-trends before election results are announced, as engagement with hate is driven up prior to polling, in the campaigning period. This however does not explain why treated villages in the sub-sample of close electoral races, were more likely to engage with toxic content

[12] `https://www.opindia.com/`

in the pre-announcement period.

We maintain that electoral races where the religious identity of the incumbent is non-Muslim and that of the probable winner is also non-Muslim, are electoral races where religious identity does not matter during the campaign period. This is because the religious identity of the candidates is not salient in such circumstances. Indeed, Figure 6 shows that toxicity scores for shared posts are the lowest in such villages. Still, the engagement patterns in all four types of villages, with respect to the polling date are indistinguishable. It is therefore, unlikely, that local campaigning activity drove up engagement with toxic content.

6 Conclusion

Using rich social media data from a very popular Indian content generation platform - Sharechat, we studied how the identity of local political leaders changes engagement with hateful content. We studied the activity of 19 million users with one million Hindi posts around the village council (*Gram Panchayat*) elections of April, 2021 in the state of Uttar Pradesh, India. To this end, we employed regression discontinuity design and off-the-shelf Hindi hate speech natural language classifiers on geolocated Panchayat election data and user data from ShareChat. We find no change in engagement with toxic content in villages where a Muslim candidate won the election. Moreover, we provide evidence that exposure to hate is highly correlated across villlages in the state of Uttar Pradesh, driven by national news stories and items by mainstream media outlets. This exposure is likely to be correlated also on account of the platform's content recommender systems that do not cuztomize content for users according to their geographic location. We discuss the production of hate content and other potential explanations for these observations.

Our present analysis suffers from the following limitations. Classification of hate speech is an inherently tricky exercise, for humans and computers alike. The performance of the hate speech classifier we used is a source of great noise in our results. A lot of work remains to be done on multilingual hate speech classifiers. At the moment, to allow a neural network to learn the political and social contexts that deem certain content hateful, one requires a huge amount of manually annotated data. Moreover, social and political contexts are ever evolving and most hate speech classifiers aren't generalizable to contexts it was not trained on. The structure of and numerous ways of spelling a word in the Hindi language also makes natural language processing tasks challenging. A noteworthy direction would be to utilize the images of posts to learn visual context clues present in hateful content.

Another limitation of our study is our sample size of villages in Uttar Pradesh. Geolocated data for rural Uttar Pradesh is sparse and ill maintained. To add strength to our methods, we hope to collect data on and analyze local political and economic shocks in the State of Bihar, India. This paper, therefore, paves way for future work on the local determinants of hate speech. This would enable our understanding of this widespread phenomenon, in order to recommend relevant policy instruments for course correction.

References

H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, 2020.

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

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

M. Das, P. Saha, B. Mathew, and A. Mukherjee. Hatecheckhin: Evaluating hindi hate speech detection models. *arXiv preprint arXiv:2205.00328*, 2022.

J. Drèze. The revolt of the upper castes. *CASTE/A Global Journal on Social Exclusion*, 1(1):229–236, 2020.

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

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

F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

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

M. Ledwich and A. Zaitsev. Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*, 2019.

T. Mandl, S. Modha, A. Kumar M, and B. R. Chakravarthi. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for information retrieval evaluation*, pages 29–32, 2020.

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

MoSPI. Village directory, population census, 2011.

R. Mukherjee. Mobile witnessing on whatsapp: Vigilante virality and the anatomy of mob lynching. *South Asian Popular Culture*, 18(1):79–101, 2020.

T. Santosh and K. Aravind. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 310–313, 2019.

A. Srivastava, A. Kalra, and S. Tiwari. A dataset of geolocated villages and gram panchayat election candidates in uttar pradesh. 2022.

Appendix A: Figures

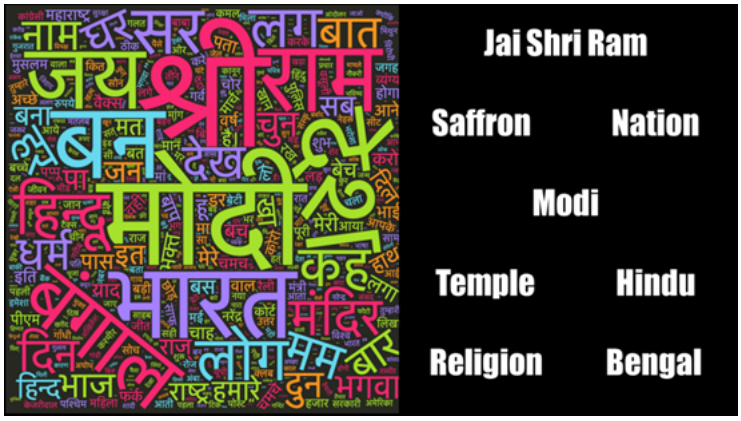


Figure A.1: Most popular phrases and words in Hindi Political posts on ShareChat, represented as a word cloud. Here weights are determined by frequency of occurrence of the text. Corresponding English translations on the right.

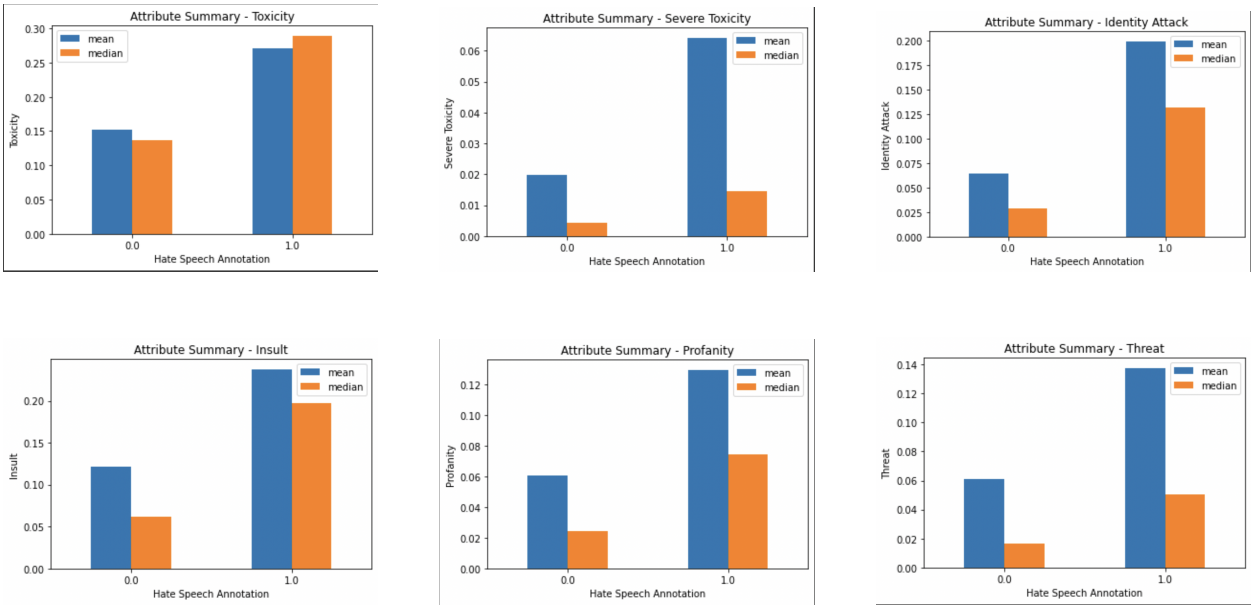


Figure A.3: Performance of Perspective AI's Toxic speech classifier wrt manually annotated ShareChat posts

Appendix B: Tables

Threshold = 0.4		
Predicted	0	1
Actual		
0	633	35
1	117	17
Threshold = 0.35		
	606	62
	105	29
Threshold = 0.32		
	592	76
	82	52

Table B.1: Columns label hateful content according to the classifier and rows label hateful content according to the manual annotation. Magnitudes display number of posts correctly classified, and number of posts classified with error.

Post OCR Text	Translated Text
गजब बात है भाई.. जमीन को खोदो तो मंदिर निकलते हैं और घोटालों को खोदो तो कांग्रेस निकलती है मनोरम SB दोनो को खोदो तो चमचे ओर जेहादीयो की जान निकलती है	It is a wonderful thing brother.. If you dig the land, then the temple emerges and if you dig the scams, then the Congress comes out.
हमें ताकतवर नहीं x दिमाग वाला प्रधानमंत्री चाहिए था। ताकतवर तो सांड भी होता है।	We wanted a prime minister with no x brains, a mighty one. Even a bull is powerful.
48% 17:17 ... ये रहा एक करोड़ का सवाल- पप्पू ज्यादा मंदबुद्धि है या उसे वोट देने वाले ??	48% 17:17 ... here is the question of one crore - is Pappu more retarded or those who vote for him??
सुअरों से सावधान रहें सु- सुधीर चौधरी अ- अर्नब गोस्वामी रो- रोहित सरदाना	Beware of pigs Su- Sudhir Chaudhary a- Arnab Goswami cry- Rohit Sardana
अधर्म का नाश हो कांग्रेस का सत्यानाश हो प्राणियों में सद्भावना हो विश्व का कल्याण हो चारो तरफ BJP का राज हो देश का विकास हो	Destruction of unrighteousness, annihilation of Congress, goodwill among the living beings, welfare of the world, the rule of BJP everywhere, development of the country.
जब तक मुख्यमंत्री नहीं बना विधानसभा में नहीं घुसा जब तक प्रधानमंत्री नहीं बना संसद में नहीं घुसा जब तक राम मंदिर नहीं बना अयोध्या में नहीं घुसा इतना जिद्दी आदमी आज तक नहीं देखा।	Till the Chief Minister did not enter the assembly, till the Prime Minister did not enter the Parliament, till the Ram temple was not built, he did not enter Ayodhya till today.
राम मंदिर बनाने से कोरोना नहीं जाने वाला - शरद पवार गल्ला L.Gossi बाला साहब ने कहा था हिजड़े वो नहीं जो सड़क पर साड़ी पहन के घूमते हैं हिजड़े वो हैं जो हिंदू होकर भी हिंदुत्व का विरोध करते हैं	Will not go to Corona by building Ram temple - Sharad Pawar Galla L.Gossi Balasaheb had said that eunuchs are not those who roam on the road wearing saris.
मुगल और अंग्रेज यहां लड़ने के लिए अपने सैनिक नहीं लाए थे क्योंकि उन्हें यहां के गद्दारो पर पूरा भरोसा था..!	Mughal and British did not bring their soldiers to fight here because they had full faith in the traitors here..!
जब डोनाल्ड ट्रम्प को 4 साल की सत्ता जाने पर इतना गुस्सा है.... तो काँग्रेस की 60 साल की सत्ता जाने पर उनकी बवासीर तो तेज होगी ही..!	When people are so angry at Donald Trump's 4 years of power.....then their hemorrhoids will only intensify if Congress goes to power for 60 years..!
चमचों तुम तो सिर्फ डंडा खाओ और झंडा उठाओ क्योंकि... जब मम्मी आउट तो बेटा बैटिंग करेगा जब बेटा आउट तो मम्मी बैटिंग करेंगी	You just eat the stick and raise the flag because... when mommy out, son will bat, when son is out, mommy will bat

Table B.2: Examples of text classified as hateful according to Perspective AI classifier (toxicity score ≥ 0.35) and manually classified as non-hostile by annotators.