Bridging or Breaking: Impact of Intergroup Interactions on Religious Polarization

Rochana Chaturvedi*
rchatu2@uic.edu
University of Illinois Chicago
Chicago, USA

Sugat Chaturvedi*
sugat.chaturvedi@ahduni.edu.in
Ahmedabad University
Ahmedabad, India

Elena Zheleva
ezheleva@uic.edu
University of Illinois Chicago
Chicago, USA

ABSTRACT

While exposure to diverse viewpoints may reduce polarization, it can also have a *backfire effect* and exacerbate polarization when the discussion is adversarial. Here, we examine the question whether intergroup interactions around important events affect polarization between majority and minority groups in social networks. We compile data on the religious identity of nearly 700,000 Indian Twitter users engaging in COVID-19-related discourse during 2020. We introduce a new measure for an individual's group conformity based on contextualized embeddings of tweet text, which helps us assess polarization between religious groups. We then use a meta-learning framework to examine heterogeneous treatment effects of intergroup interactions on an individual's group conformity in the light of communal, political, and socio-economic events. We find that for political and social events, intergroup interactions reduce polarization. This decline is weaker for individuals at the extreme who already exhibit high conformity to their group. In contrast, during communal events, intergroup interactions can increase group conformity. Finally, we decompose the differential effects across religious groups in terms of emotions and topics of discussion. The results show that the dynamics of religious polarization are sensitive to the context and have important implications for understanding the role of intergroup interactions.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Social and professional topics** → **Religious orientation**; • **Applied computing** → **Economics**; **Sociology**.

KEYWORDS

Social media, Polarization, Intergroup Interaction, Religion

ACM Reference Format:
Rochana Chaturvedi, Sugat Chaturvedi, and Elena Zheleva. 2024. Bridging or Breaking: Impact of Intergroup Interactions on Religious Polarization. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589334.3645675

*Both authors contributed equally to this work.

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05
https://doi.org/10.1145/3589334.3645675

1 INTRODUCTION

Polarization between identity groups drives social unrest and adversely affects a nation's economic growth and responses to crises [21, 22]. However, it is less clear how polarization evolves during crises. While collective suffering may foster within-group solidarity [7], it may also lead to attribution of blame on the "outside" group or increase between-group competition for limited resources [32]. These behaviors might be accentuated among people who interact only within their groups and hence might have a restricted information environment [31]. Polarization has been widely studied in the social-media context as well. In particular, social media platforms such as Twitter have become more polarized over time [26]. Consistent with this, political polarization decreased for users who deactivated their Facebook accounts before the US elections [1]. The depolarization, however, depends on an individual's background and this pattern might reverse for individuals having homogeneous offline social networks [4]. Interactions within social media platforms might have varying impacts on polarization. In an unfavourable environment, exposure to outgroup viewpoints may lead to stereotype formation and increase polarization. For example, Bail et al. [5] conduct a field experiment and find that Republicans (Democrats) who were offered financial incentives to follow a liberal (conservative) Twitter bot became more conservative (liberal).

In this paper, we examine the impact of intergroup interactions on polarization on Indian Twitter between religious majority and minority groups in the context of COVID-19-related events. We consider an individual as engaging in intergroup interaction if they post a reply to someone outside their own group. We introduce a new measure of an individual's conformity to their group based on contextualized embeddings of tweet texts. We call this a user's **Group Conformity Score (GCS)** which measures how similar the user's tweets are to their own group as opposed to tweets by users of the other group. Polarization is the sum of GCS over all the users, weighted by the inverse of their group size. We then unveil heterogeneous effects of intergroup interactions on a person's group conformity over different pandemic-related events using a meta-learning framework. We examine the heterogeneities in the effect across religion, topics, emotions, engagement, and ego-network. Finally, we decompose the differences between the treatment effects for the two religious groups into the treatment effects on change in topics of discussion and on change in emotions.

The standard measure of ideological polarization DW-NOMINATE proposed by Poole and Rosenthal [41] defines polarization in terms of the distance between Republicans and Democrats in the US legislature based on roll call voting. Gentzkow et al. [27] extend this by proposing a bag-of-words (BOW)-based estimator to measure partisanship in congressional speech and find a correlation of

53.7% with DW-NOMINATE. They argue "If two (political) parties speak more differently today than in the past, these divisions could be contributing to deeper polarization in Congress." Additionally, speech may be shaped by fewer strategic considerations compared to roll-call voting and can be used to estimate polarization across more diverse contexts. Their estimator overcomes finite sample bias resulting from phrases that a group might simply mention by chance. Demszky et al. [18] use this estimator to examine polarization on Twitter in the context of 21 mass shooting incidents in the US, demonstrating the link between the polarization measure on social media and real-world events. They argue that there are many ways in which polarization can be instantiated linguistically and interpret this measure in terms of topic choice, framing, affect, and illocutionary force. They find that polarization, as captured by this measure, is driven by partisan differences in framing rather than simply topic choice. Further, they find that controlling for the total number of followed politicians, an additional followed politician from one's preferred party is associated with an increase of .009 SD in the polarization measure. In contrast, controlling for the total number of followed politicians from the user's preferred party, an additional followed politician is associated with a decrease of .02 SD in the polarization measure. Therefore, their work contributes to validating the linguistic polarization measure even in the less formal social media setting. One limitation of the bag-of-words representation is that it does not take the context or synonymy in two phrases into account. Our new contextualized-embeddings-based measure (GCS) addresses this by capturing different dimensions of *linguistic polarization*, conceptualized to exist when the two groups semantically diverge in their tweets, more meaningfully.

Several studies identify ideological polarization on social networks based on interaction network clusters [15] or concentration of influential users at the boundaries of two communities [29]. Others detect polarized communities based on whether the interactions between users are friendly (positive) or antagonistic (negative) [11]. We abstract away from the problem of positive or negative interactions in network-based metrics by measuring polarization using the content produced by well-identified groups. One way to define polarization is by modeling the propagation of users' opinions and the differences in opinions at the equilibrium [38, 39]. Another way is to characterize *affective* polarization based on toxicity in intergroup interactions [20]. Garimella et al. [25] use graph clustering and network-based metrics to identify polarizing topics. Other measures incorporate only specific dimensions of tweet content such as stance [17] or contextualized embeddings of keywords [19].

Our work is motivated by the contact hypothesis that intergroup contact can reduce prejudice towards the outgroup when groups engage in equal status contact in the pursuit of common goals and in the presence of intergroup cooperation under a favorable institutional environment [3]. Thus, intergroup interaction can lead to a better understanding of outgroup perspectives and lead to cross-group friendships [12, 43]. At a broader level, this may facilitate national integration but may have the opposite effect in polarized settings [8]. In the political arena, Levendusky and Stecula [34] experimentally demonstrate that cross-party discussions decrease affective polarization (or dislike of outgroup individuals) between Republicans and Democrats in the US. This decrease is conditional on conversation topics not involving disagreements [46]. The effects of intergroup contact might vary for majority and minority groups with possibly weaker effects for minorities [47]. Interestingly, intergroup contact focusing on commonalities between majority and minority groups can lead the minority to perceive the majority group as fairer than they are [45]. In the Indian context, Lowe [35] randomly assign individuals from different caste groups to the same (collaborative contact) or opposing (adversarial contact) cricket teams. They find that collaborative contact increases cross-caste friendships while adversarial contact has the opposite effect—thus highlighting the importance of the setting. In the online context, intergroup conversations between Hindus and Muslims on Whatsapp are found to decrease prejudice against Muslims [36]. We add to this evidence by going beyond prejudice and focus on group conformity in tweet text for both majority and minority groups and examine the heterogeneous effects of intergroup contact. Given the underlying tensions between Hindus and Muslims, often resulting in large-scale violence, detecting religious polarization between them on social media is of particular relevance.

In line with the above discussion, we hypothesize that in general intergroup interaction should decrease Group Conformity Score (GCS), and thus polarization. However, such interaction should be less likely to decrease GCS for individuals with already entrenched positions and who might be less receptive to outgroup perspectives. Further, when individuals in the minority group are disproportionately affected by an event, we expect intergroup interaction to amplify GCS for them. We expect the opposite effect for the unaffected majority group who might become sympathetic to minority issues due to interaction. Finally, for politically salient events, intergroup interaction should increase polarization for individuals having a high predisposition towards political discussions and who might have conflicting ideologies.

2 DATA

2.1 COVID-19 Tweets India

We use the "Global Reactions to COVID-19 on Twitter" data collected by Gupta et al. [30]. The core data comprise over 132 million English language tweets from more than 20 million unique users using 4 keywords—"corona", "wuhan", "nCov", and "COVID". The tweets were posted during January 28, 2020–January 1, 2021.[1] We use the India sample of the data, i.e. tweets about or originating from India.[2] Hydrator application is used to obtain complete information on tweets from their IDs (collected on May 4, 2021). Out of a total of 6,166,152 tweet IDs, full data for 5,459,402 tweets could be collected representing an attrition rate of 11.46%. This is due to the deletion of some of the tweets and accounts by the collection date. These tweets are by 871,203 unique users. The tweets are cleaned by removing mentions, hyperlinks, and extra whitespaces. The data contains information on the user name, their account creation date, number of friends and followers, and whether a tweet is a retweet

[1]The dataset of tweet IDs is publicly available at https://doi.org/10.3886/E120321V6.
[2]This restriction is imposed by mapping the user location attribute on Twitter to country using GeoNames' cities15000 geographic database available at http://download.geonames.org/export/dump/cities15000.zip. The place field and the user location field are mapped to India by matching them with a dictionary of cities and states in India. For places that could not be mapped to India using the previous step, we use Nominatim—a search engine used for OpenStreetMap (OSM). This gives the complete address of a place and allows us to remove tweets posted from outside India.

or a reply. It also contains information on five psycho-linguistic attributes for each tweet indicating the intensity of valence, anger, fear, sadness, and joy extracted using CrystalFeel—"a collection of machine learning-based emotion analysis algorithms for analyzing the emotional-level content from natural language".[3]

2.2 Events

We adopt a principled approach to compile a comprehensive list of significant events in India in 2020. We begin by consulting reputed sources such as major daily news outlets and well-curated information from Wikipedia. This ensures capturing a broad spectrum of perspectives and including widely acknowledged and reported events. Our inclusion criteria focus on events that substantially impacted Indian society, politics, economy, or culture. This approach combining external sources with internal expertise enables us to filter out events that were potentially newsworthy but did not meet the threshold for major and impactful occurrence. For each event, we consider the subset of tweets seven days post-event (inclusive of event date) period. We count the number of tweets that contain event-related keywords within this subset. Since our dataset comprises only COVID-19-related tweets, this count gives us the importance of the event within the context of pandemic-related conversations—ensuring a well-rounded and contextually relevant selection. We provide the list of all events, event-specific regular expressions, and the number of tweets matching the regular expression in Appendix S1 Table S1. We identify seven highly discussed events based on the tweet counts and describe them in detail in Table 1. Our subsequent analysis is based on these seven events.

3 METHODOLOGY

To study the treatment effect of Intergroup Interactions on Group Conformity of an individual, we first augment our dataset to include all the variables of interest. We describe these in Sections 3.1–3.3. We then describe the steps for treatment effect estimation in Section 3.4 and the decomposition of differences in treatment effects across religions in terms of topics and emotions in Section 3.5.

3.1 Inferring Religion

Since the data on religious identity is not available on Twitter, we infer this from usernames. In India, names are highly predictive of group identity and we leverage character sequence-based machine learning models from our earlier work [13] to obtain religion estimates.[4] Previous research uses these to examine disparities in allocation of publicly provided goods [14], impact of government surveillance on minority voter turnout [2], and potential election irregularities [16]. We classify the religion of each user as Muslim (also referred to as the minority as Muslims are the largest religious minority in India and comprise 14% of the total population according to Census 2011) or non-Muslim (alternatively referred to as Hindus or majority group comprising 80% of the total population).[5] It is worth discussing the potential ethical implications of inferring religion or group identity from names in the social media context. Though such an algorithm can be used by nefarious actors, it can also help researchers highlight systematic targeted harassment or, as in our case, monitor trends in polarization across identity groups. See [13] for a more general discussion on ethical concerns.

We drop verified users from our data to remove influential individuals/organizations. To reduce possible biases in the polarization measure due to erroneous classification of religion, we remove non-personal names. We discuss the details of name cleaning steps in Appendix A. Our final data comprises 692,559 unique users with 3,072,503 tweets.[6] The model from [13] is trained on names obtained from a nationally representative sample, while Twitter usernames tend to be noisy. The name classification exercise also depends on the distribution of names in the specific domain the algorithm was trained on. Therefore, we expect a domain shift when applying the model to our data. To address this, we manually annotate a subset of approximately one thousand names as Muslim or non-Muslim. Since Muslim and non-Muslim names are linguistically distinct, prior literature has also relied on manually annotating religion as Muslim and non-Muslim from names in the South Asian context. In Appendix B, we discuss how we select the sample of names for manual annotation and choose a more suitable classification threshold. Chaturvedi and Chaturvedi [13] report an F_1 score of 95% on their test set which reduces to around 85% when only one name part is available (i.e., only the first or last name)—as is often the case on Twitter; we find comparable values (sensitivity 84% and specificity 86.5%). The misclassifications lead to measurement error, potentially underestimating polarization; though qualitatively we expect the temporal patterns to remain the same. Therefore, our estimates are likely to be conservative estimates. Similarly, since we focus on changes in GCS over time to estimate the treatment effects (see Section 3.4), our results should qualitatively hold.

3.2 Measuring Polarization

We use the estimated religious identities to measure polarization in terms of conformity of each user to their religious group.

3.2.1 ***Polarization via Bag-of-Words.*** We first consider the leave-out estimator of phrase partisanship proposed by Gentzkow et al. [27].[7] We compute daily user-level polarization or the bag-of-words-based Group Conformity Score ($GCS^{BOW}_{i,d}$) using the same implementation as in [18] as the following dot product:

$$GCS^{BOW}_{i,d} = \hat{q}_{i,d} \cdot \hat{\rho}_{-i,d}$$

Where $\hat{q}_{i,d}$ is the vector of token (unigram and bigram) frequencies (c_i) normalized by the sum of all token counts (m_i) for user i on day d, only considering tokens used by at least two tweeters. $\hat{\rho}_{-i,d}$ is the vector denoting the sum of normalized token frequencies across all users in i's group $g_i \in \{Muslim(M), non-Muslim(NM)\}$ while leaving i out, relative to users in the other group $\tilde{g}_i$.

$$\hat{\rho}_{-i,d} = \hat{q}_d^{g_i - \{i\}} \oslash (\hat{q}_d^{g_i - \{i\}} + \hat{q}_d^{\tilde{g}_i})$$

[3] See https://socialanalyticsplus.net/crystalfeel/.

[4] We use the Single Name SVM model to obtain the Muslim score as recommended.

[5] Other religious groups comprise Christians (2.3%), Sikhs (1.7%), Buddhists (0.7%), and Jains (0.4%). Therefore, 93% of non-Muslims are Hindus.

[6] To check for the possibility that our results might be influenced by bots, we use the recent lists TwiBot-20 [24] and TwiBot-22 [23]. We find that less than 0.5% of users in our data are listed as bots and contribute to 0.63% of the tweets.

[7] Before applying this estimator, we lower-case the tweets, remove stopwords (see Appendix S2 for the stopwords list) and punctuations, and stem words using the NLTK's Snowball Stemmer. Removing stopwords leads to dropping 1,007 user-day observations comprised entirely of stopwords and punctuations.

Table 1: Description of COVID-related events discussed highly in COVID Tweets India subset in year 2020

Event	Date	Description
Janata Curfew	Mar 22	A day-long curfew announced by the government for all citizens barring essential services to curb the pandemic.
Tablighi	Mar 31	Tablighi Jamaat, a Muslim congregation in Delhi, defied a ban on public gatherings during the pandemic, leading to a COVID-19 super-spreader event. Reports of attendees spitting on doctors and a viral hashtag #CoronaJihad fueled Islamophobia. The Supreme Court later criticized media for communalizing the incident.
Migrant Deaths	May 8	An estimated 10 million workers were forced to undertake long arduous journeys back home on foot after losing jobs due to abrupt pandemic-related lockdown and suspension of train services. 16 of them were killed by an empty goods train while they were sleeping on the tracks on this day.
Coronil Launch	Jun 22	Indian multinational conglomerate Patanjali spearheaded by popular yoga guru Ramdev launched an ayurvedic remedy claiming to cure COVID-19. It was approved by the Ministry of Ayush (for traditional medicine) even though there was no clinical data to support the claim. This led to controversy on social media—massive praise from some and harsh criticism from others. After the controversy, sales were halted but later permitted as an immunity booster. Some state governments, deeming it a fake medicine, imposed a complete ban.
Exam Satyagraha	Aug 23	All India Students' Association organized one-day hunger strike and satyagraha against the government's in-person national-level exams citing health risks due to COVID-19, logistical challenges from lockdowns, and suspension of public transport. More than 4000 students participated and multitudes showed support via social media.
GDP Contraction	Aug 31	Indian government announced the biggest economic slump in GDP that India had seen in 24 years.
BJP Bihar Manifesto	Oct 22	The ruling political party (BJP) promised free vaccines for all in Bihar during the Assembly election sparking criticism across religious groups over social media with the hashtag #VaccineForVotes.

Here, $\oslash$ indicates element-wise division and $\hat{q}_d^g = \sum_{j \in g} \hat{c}_j / \sum_{j \in g} \hat{m}_j$. This can be interpreted as the posterior probability that an observer with a neutral prior would assign a tweeter their true group identity after observing a single token drawn randomly from their tweets, though this interpretation relies on the assumption that a user's phrase choice is independent of other phrases used by them. Intuitively, $GCS_{i,d}^{BOW}$ captures similarity in phrase usage for user i with their group members relative to the similarity with the other group. The daily polarization is then estimated as the following average:

$$\hat{\pi}_d^{LO,BOW} = \frac{1}{2} \sum_{g \in \{M,NM\}} \frac{1}{|g|} \sum_{i \in g} GCS_{i,d}^{BOW}$$

3.2.2 ***Polarization via Contextualized Embeddings.*** The bag-of-words-based polarization estimator ignores the larger context of a tweet and can have several limitations. Firstly, distinct words (for example, greetings such as *salaam* vs. *namaste*) used by two groups conveying the same underlying message will contribute positively towards the polarization estimate. Secondly, if two users have different stance on a given issue while having broadly similar phrase usage (for example, *Coronil cures Covid* vs. *Coronil does not cure Covid*), the BOW estimator will consider them to be similar.

We address these by computing the contextualized-GCS score or simply $GCS_{i,d}$ for user i on the day d to estimate the measure at the daily level. For this, we map all the tweets to a 768-dimensional vector space using a sentence-transformer—specifically, the pre-trained *all-mpnet-base-v2* model [44]. This model is fine-tuned on over a billion sentence pairs from diverse domains and has shown state-of-the-art results on semantic search and sentence embedding tasks.[8] The measure *GCS* does not require the assumption of independence among a user's phrase choice and remains tractable due to the computational efficiency of sentence transformers. We first average these embeddings at the user-day level $u_{i,d}$ and then compute daily centroids for both the groups by taking the mean of $u_{i,d}$ across all users in a group. Averaging a user's tweet embeddings allows us to obtain a single representation for a group of sentences to collectively model a user's opinions on a given day. This ensures that the daily group centroids are not biased towards users with higher tweet frequencies. Analogous to the leave-out estimator, we first adjust the group centroid by subtracting $u_{i,d}$ from it before computing $GCS_{i,d}$. Thereafter, we compute the distances between $u_{i,d}$ and both the centroids. Finally, $GCS_{i,d}$ is computed as the Euclidean distance from the other group's centroid relative to their own adjusted centroid.[9] We use the following formula:

$$GCS_{i,d} = \frac{\|u_{i,d} - \frac{1}{|\tilde{g}_i|} \sum_{j \in \tilde{g}_i} u_{j,d}\|}{\|u_{i,d} - \frac{1}{|g_i - \{i\}|} \sum_{j \in g_i - \{i\}} u_{j,d}\| + \|u_{i,d} - \frac{1}{|\tilde{g}_i|} \sum_{j \in \tilde{g}_i} u_{j,d}\|}$$

Higher values of $GCS_{i,d}$ correspond to greater conformity of a user to their group. The daily polarization $\hat{\pi}_d^{LO}$ is computed by aggregating this measure across users in the two groups as before:

$$\hat{\pi}_d^{LO} = \frac{1}{2} \sum_{g \in \{M,NM\}} \frac{1}{|g|} \sum_{i \in g} GCS_{i,d}$$

3.3 Discussion Topics During COVID-19

To identify major topics discussed around COVID-19 events and to include them as covariates for examining the effect of intergroup interaction on change in *GCS*, we perform topic modeling over the tweets. We leverage contextualized embeddings for this task as well, following the approach of Grootendorst [28] who use sentence transformers to obtain document embeddings before using a clustering algorithm. We use the subset of tweets considered for treatment effect estimation across all the events and drop duplicate tweets so that retweeting does not affect topic assignment. We then cluster the tweets' contextual embeddings obtained using sentence transformer model *all-mpnet-base-v2* using k-means clustering algorithm.[10] To infer representative topic labels, we preprocess the

[8] It is openly and freely available and provides consistent embeddings. For more information, see https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[9] The GCS^{BOW} formula is defined in terms of similarity to one's group, while for *GCS* we consider the distance from the other group. Therefore, the broad operationalization of the two metrics is consistent.

[10] We get 7 as the optimal cluster number based on manual scanning and elbow method heuristic by plotting inertia against the number of topics over 3 to 10 topic clusters.

tweets by first lower-casing them. Since the entire dataset comprises COVID-specific tweets, we remove COVID and its synonyms for a more meaningful inference of topic labels. We also replace different vaccine names with the word vaccine, remove mentions, URLs, numbers, special HTML entities such as "&amp" and "&quot", punctuation, and extra spaces. We then transform each tweet by joining together commonly occurring multi-word phrases in that tweet using the Gensim phrase model [37]. We then concatenate all the tweets within a topic as a single document. Finally, we compute class-based TF-IDF defined as:

$$cTF\text{-}IDF_i = \frac{t_i}{w_i} \cdot log\frac{m}{\sum_{j=1}^{n} t_j}$$

Where t_i is the frequency of a word/phrase within the i^{th} topic and w_i is the total number of phrases in the topic. The total number of tweets (or unjoined documents) is m and is normalized by the number of occurrences of the word/phrase across all n topic clusters.

We identify the following topics: General COVID response, COVID prevention, COVID News/statistics (general), COVID news/statistics (state-specific), Socio-Economic, Political-Religious, and China & Global. We do a qualitative and quantitative analysis of the topic clusters and find that the COVID-specific topics are very similar and merge them into a single topic COVID Response to get the final four topics.[11] We provide the fifty most representative phrases associated with each of these topics and a sample of tweets associated with each topic in Appendix S3 Tables S2 and S3, respectively.

3.4 Conditional Average Treatment Effect

In this section, we describe our methodology to answer the question *do intergroup interactions change a user's conformity to their group?* We first describe the treatment and outcome variables:

3.4.1 ***Treatment: Intergroup Interaction.*** For each event, we look at all tweets before the event. We consider tweets that are replies and check the religion of the user posting the reply and that of the user being replied to. Our treatment variable *interact* is a binary indicator that equals 1 if a user replies to someone outside their group at least once, and 0 otherwise.[12]

3.4.2 ***Outcome.***

Change in GCS. For each event, we define an event window of n days before and after it. We compute the n-day mean of GCS_i for each user i over the pre-event and post-event windows.[13] Finally, we take $\Delta GCS_i = \overline{GCS}_{i,post} - \overline{GCS}_{i,pre}$ as the difference in the averages post and pre-event. We choose the window size to be large enough to balance the daily fluctuations and small enough to rule out other events influencing the outcome.

Change in Topics and Emotions. We also estimate the effect of intergroup interaction on changes in topics and emotions. This helps decompose the differential effects of intergroup interaction on ΔGCS_i in terms of differential effects on changes in topics and emotions across religions. We again take the mean difference in these variables across post and pre-event windows for each user.

3.4.3 ***Pre-treatment Covariates.*** We consider 30 days pre-event window and compute the following covariates for adjustment: 30-day averages of *GCS*, emotion intensities for valence, anger, fear, sadness, and joy; ego-network features such as friends and followers counts; engagement features such as tweet frequency, average number of times a user's tweets were retweeted and the fraction of replies among tweets; the number of days lapsed since account creation to event date; Muslim score as given by the religion classifier; and lastly the fraction of user tweets in the pre-treatment period assigned to each topic. We use inverse hyperbolic sine transformation (arcsinh) for friends and followers counts, tweet frequency, and average retweets, as these are right-skewed. Arcsinh approximates logarithmic transformation while allowing us to retain zero values. We then normalize all the covariates and the outcome variable.

The descriptive statistics for the final event-level dataset are provided in Appendix D Table 4.

3.4.4 ***T-Learner.*** Metalearners [33] combine predictions from standard supervised machine learning algorithms to estimate heterogeneous treatment effects through the Conditional Average Treatment Effect (CATE) estimand defined as:

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid X = x] = M_1(X) - M_0(X)$$

Where $Y(1)$ is the potential outcome for an individual if they were treated ($T = 1$), i.e. if they interacted outside their group; $Y(0)$ is the potential outcome if the same individual belonged to the control group and were not treated ($T = 0$); X is the vector of pre-treatment covariates. We focus on ΔGCS, the change in Group Conformity Score, as our main outcome. We observe either $Y(1)$ or $Y(0)$ for any given individual, and use T-Learner [33] to estimate the unobserved potential outcome and then CATE in the following stages:

(1) **Training Stage**: In the first stage, we estimate conditional expectations of the outcomes $\hat{M}_1$ and $\hat{M}_0$ using observations from the treatment and control groups, respectively.
(2) **Prediction Stage**: We then estimate Individual Treatment Effect ITE for i^{th} user using predictions from $\hat{M}_1$ and $\hat{M}_0$ over the complete set of observations in the test set as:

$$\hat{\tau}(x_i) = \hat{M}_1(x_i) - \hat{M}_0(x_i)$$

Estimation of τ requires assuming ignorability or that there are no unobserved confounders, i.e. covariates jointly influencing the treatment and the outcome. We control for pretreatment *GCS* which can encapsulate information on an individual's prior exposure to the outside group or other unobservable factors that may affect the outcomes. We also control for pretreatment covariates such as topics, emotions, ego-network, and engagement features as discussed in Section 3.4.3 to account for other individual-specific factors.

Implementation Details. We compile event-specific subsets by considering the outcomes ΔGCS and changes in topics and emotions within a 7-day window post and pre-event. We combine this with the treatment variable and pre-event covariates (see Section 3.4.3). To estimate the response functions $\hat{M}_1$ and $\hat{M}_0$, we use nested Lasso with 10-fold cross-validation (CV). For $\hat{M}_1$, we use the subset of data with $T = 1$ and for $\hat{M}_0$ that with $T = 0$. We split the subsets

[11] The qualitative analysis considers the most frequent words associated with each topic while quantitative analysis examines distance between cluster means and mean of pairwise cosine distances across clusters. Our results for metalearners in Section 4.2 remain the same with and without merging topic clusters.
[12] The users who never reply to anyone are dropped from further analysis
[13] Users who do not tweet during any window are dropped.

into outer 10 folds, and for each iteration of the outer 10-fold CV, we further split the training fold into inner 10 folds for hyperparameter tuning. We use the best model from the inner 10-fold CV to estimate the outcomes and evaluate on the outer fold.[14]

3.5 Oaxaca-Blinder Decomposition

Finally, we decompose the mean of $\Delta\hat{\tau} = \hat{\tau}_{Muslim} - \hat{\tau}_{non-Muslim}$ into effects on each topic and emotion. We use the Oaxaca-Blinder method [10, 40] to decompose the differences at the *mean* into *explained* and *unexplained* components and further into contributions of individual covariates to explained differences.

Given $\bar{\hat{\tau}}_g = \sum_x \beta_g^x \bar{\hat{\tau}}_g^x$, $g \in \{Muslim(M), non-Muslim(NM)\}$, where β_g^x are the regression coefficients and $\bar{\hat{\tau}}_g^x$ are mean values of covariates (the average treatment effects on $x \in topics \vee emotions$):

$$\begin{aligned}\Delta\bar{\hat{\tau}} &= \bar{\hat{\tau}}_M - \bar{\hat{\tau}}_{NM} \\ &= \sum_{x \in topics \vee emotions} \beta_M^x \bar{\hat{\tau}}_M^x - \beta_{NM}^x \bar{\hat{\tau}}_{NM}^x \\ &= \sum_{x \in topics \vee emotions} \underbrace{\beta_{NM}^x (\bar{\hat{\tau}}_M^x - \bar{\hat{\tau}}_{NM}^x)}_{\text{Explained}} + \underbrace{\bar{\hat{\tau}}_M^x (\beta_M^x - \beta_{NM}^x)}_{\text{Unexplained}}\end{aligned}$$

The explained component captures what part of the mean difference in treatment effect across Muslims and non-Muslims is due to differential effects on topics and emotions. The residual or unexplained component captures to what extent the marginal effect of each covariate on the outcome is different across the two groups, given that they have the same explanatory attributes.

4 RESULTS

Here, we first discuss a qualitative analysis based on our proposed metric *GCS* in Section 4.1 and then the results from treatment effect estimation in Section 4.2. Finally, we discuss the decomposition of differences in the treatment effects across the two religious groups into the effects on changes in topics and emotions in Section 4.3.

4.1 Qualitative Content Analysis based on GCS

To compare the BOW and contextualized-embeddings-based estimators, Figure 1 plots the seven-day exponential moving average of polarization trends using $\hat{\pi}^{LO,BOW}$ and $\hat{\pi}^{LO}$. We find similar trends in the daily aggregate polarization values using both the measures with Pearson's correlation of 66.42% (significantly different from 0; p-value = 0.000). However, the fluctuations in BOW polarization are more pronounced. The polarization increased during the Tablighi incident on March 31, 2020, which was marked by increased Islamophobic sentiments. The highest peaks are during the Muslim festivals—the beginning of the holy month of Ramadan and its culmination in Eid-ul-fitr. There are also smaller peaks during the Muslim festivals of Eid-ul-Zuha and Eid-e-Milad. We find that the Muslim tweets during these festivals are mostly greetings and well wishes while non-Muslim tweets discuss a variety of subjects. We do not find any peaks around non-Muslim festivals.

[14]Base learners such as support vector regression (SVR), random forest, Ridge, or RANSAC regression with grid-search for hyperparameter tuning lead to worse MSE and R-square. We also experiment with window sizes of 5 or 10 days and find broadly similar trends. We get qualitatively similar results using X-learner [33].

To understand whether *GCS* provides a more meaningful measure compared to GCS^{BOW}, we examine tweets by the top thirty high-*GCS* users in both religious groups across the 7-day window post each event. All *GCS* values in this subset lie above the 98^{th} percentile for each event-religion-GCS (i.e., *GCS* or GCS^{BOW}) combination. We present five example tweets from these in Appendix S4 Tables S4 (high *GCS*) and S5 (high GCS^{BOW}). As expected, the tweets from users in one group having high group conformity are similar to tweets from users in the other group having low group conformity. This holds even without exactly matching but semantically similar tweets in the case of *GCS*. The highest GCS^{BOW} tweets for the majority group often comprise a few common words. This is consistent with the results in Appendix C, Figure 4 in which we examine the relation of tweet length with GCS^{BOW} and *GCS*. We find that the average tweet length is low at extreme values of GCS^{BOW} while this relation is weak in the case of *GCS*.

We now qualitatively discuss each group's high *GCS* tweets based on Appendix S4, Table S4. During the pre-lockdown **Janata Curfew**, we notice hostile attitude towards China and appreciation for frontline workers among non-Muslims. Muslims predominantly share news related to Kashmir (a Muslim-majority state) and Muslim-majority countries. After **Tablighi** incidence, non-Muslims promoted the Indian prime minister's plea to light candles in a show of unity in the fight against COVID-19 at 9 p.m. for 9 minutes, whereas Muslims expressed anger over Islamophobia spread by Indian media after this event. A noteworthy example, in the aftermath of the Tablighi incident, is that of a non-Muslim user supporting Muslims: *"The manner in which media showed propaganda against tabliqi jamat & corona jihad etc but didn't shown Bombay HC judgements which said tablighi's were made scapegoats, same way they will show propaganda against @Tweet2Rhea & @deepikapadukone but will not show u the judgments later". GCS* (0.497 or 8^{th} percentile) correctly identifies low group conformity for this user while GCS^{BOW} (0.84 or 99^{th} percentile) fails to do so.

After **Migrant Deaths** there's some discussion on the plight of migrants among high *GCS* non-Muslims. It captures two viewpoints—(i) the suffering of migrants, and (ii) increasing risk of COVID spread and economic issues resulting from migration. On the other hand, high *GCS* Muslims express anger against the government and media for not covering the issue. After **Coronil Launch**, non-Muslims express relief against COVID and pride in the Indian Ayurvedic medicine, though a few express skepticism as well. On the other hand, several high *GCS* Muslims label it a fake drug. In addition, Muslim discourse remains centered around Islamophobia and the bigotry of news media in coverage of Tablighi vs. government approval of Rath Yatra—a Hindu religious congregation.

Post the call for **Exam Satyagraha**, high *GCS* non-Muslim discourse is more varied with some concerns regarding increased COVID risk due to in-person examination while high *GCS* Muslims harshly criticize the decision of in-person exams. This is perhaps due to Muslims being the poorest religious group in India and might find it logistically harder to attend in-person exams. Similarly, during **GDP Contraction** high *GCS* non-Muslim discourse remained more varied while Muslim tweets express criticism towards the government over GDP decline. Finally, after the release of **Bihar Manifesto** by the ruling party BJP, high *GCS* non-Muslim discourse

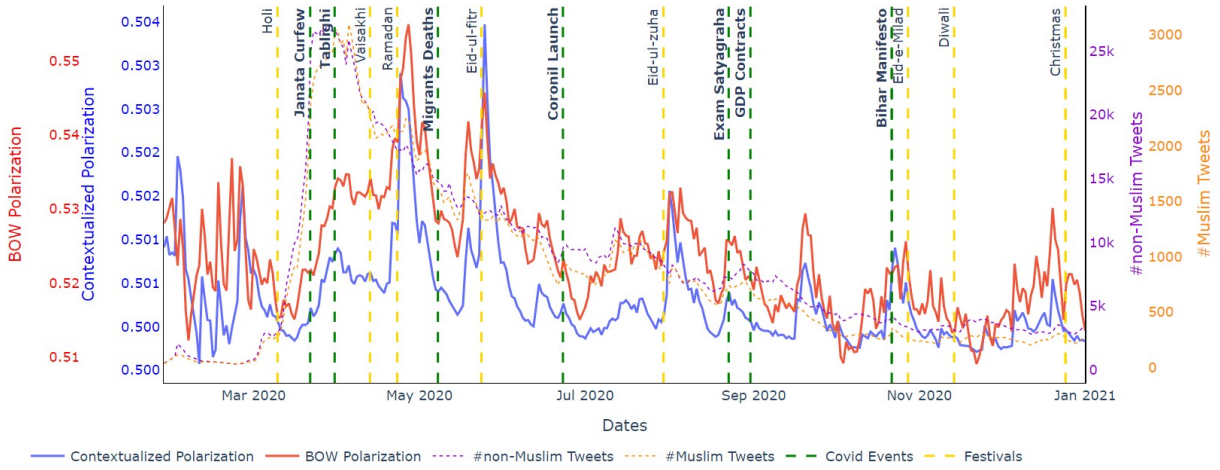


Figure 1: 7-day exponential moving average of daily polarization estimated using contextualized approach $\hat{\pi}^{LO}$ vs. bag-of-words approach $\hat{\pi}^{LO,BOW}$ along with the number of COVID-related tweets by both religious groups. The COVID-related events are marked with green vertical lines and major festivals are marked with yellow vertical lines.

focuses on general COVID-related news with some criticizing the *vaccine for vote* clause in the manifesto. This criticism appears to be unanimous among the high *GCS* Muslim tweeters.

4.2 Effect of Interaction on Change in GCS

In this section, we examine the results from T-learner. Table 2 shows the average effect $\hat{\tau}_{All}$ of intergroup interaction on change in overall *GCS* and also separately for Muslims $\hat{\tau}_{Muslim}$ and non-Muslims $\hat{\tau}_{non-Muslim}$ across all the events. Appendix S5, Figure S1 shows distributions of individual treatment effects for Muslims and non-Muslims. We find that intergroup interaction decreases overall *GCS* (or $\hat{\tau}_{All} < 0$) for all events except GDP Contraction for which there is an increase in *GCS* ($\hat{\tau}_{All} = 0.05$ standard deviations or s.d.). In other words, talking to people from the other group generally contributes to a decrease in polarization. The strongest negative effect among these is for the Tablighi incident (-0.16 s.d.). In contrast, while intergroup interaction decreases the average *GCS* for Muslims after all the other events, the effect is positive for the Tablighi event (0.04 s.d.) which was a highly communal event followed by increasing islamophobia in India. This suggests that intergroup interaction amplifies the polarizing effect of such events for the affected minorities. Notably, the negative effect for Muslims after GDP Contraction is not statistically significant while all the other coefficients we discuss are statistically significant at the 1% level of significance. The strongest negative effects of intergroup interaction on *GCS* for Muslims are in the case of Janata Curfew (-0.24 s.d.) and after the release of Bihar Manifesto (-0.18 s.d.).

We examine the heterogeneity in the treatment effect (TE) by regressing the treatment effect $\hat{\tau}$ on standardized pre-treatment covariates for each event. Appendix E, Figure 5 reports the complete results; we highlight the most important findings here. We find a high positive correlation between pre-treatment *GCS* and $\hat{\tau}$. In other words, the decline in *GCS* due to intergroup interaction is stronger for people with an already low *GCS*. This is especially true in case of the Bihar Manifesto which is a highly political event, and on which Hindus and Muslims might have divergent perspectives. However, this positive correlation breaks down for the Tablighi event. Among the topics, $\hat{\tau}$ is more negative after the launch of Coronil remedy for people who were initially more engaged in *COVID response* discussion, and hence, might have shared concerns related to this. The other notable topic is Politics-Religion with which $\hat{\tau}$ has a highly positive correlation in the case of Exam Satyagraha and Bihar Manifesto both of which are political events. Specifically, Exam Satyagraha was called for by the left-wing student organization AISA when the right-wing ruling government announced the decision to conduct exams. Among the emotions, we find a high positive correlation of $\hat{\tau}$ with Anger in case of the communally charged Tablighi event indicating that the intergroup interaction increased *GCS* for people who expressed more anger earlier.

4.3 Decomposition Analysis

Given that the effects of interaction on change in *GCS* exhibit substantial heterogeneity across the two religious groups, a natural question is—*what is the contribution of topics and emotions towards explaining these differences?* Importantly, emotions and topics are also computed as properties of the tweet text, and changes in *GCS* partially embody changes in these attributes.[15]

Figure 2 shows the aggregate decomposition into the explained and unexplained components. We observe the largest negative $\Delta\bar{\hat{\tau}}$ in case of Janata Curfew (-0.2 s.d), and Bihar Manifesto (-0.1 s.d.) and the explained component of these differences are estimated at 46.4% and 65.7% respectively. We also find a highly positive difference (0.2 s.d.) in the case of the Tablighi incident and the explained component of this difference is 7.9%. The difference in the case of GDP Contraction is (-0.07 s.d.). However, the explained component of this difference is not significantly different from 0. We also observe small negative $\Delta\bar{\hat{\tau}}$ for Coronil Launch (-0.03 s.d.) and Exam Satyagraha (0.03 s.d.). For Coronil Launch, the covariates overexplain the difference with the explained component at 147%, while for Exam Satyagraha the explained difference is 84.7%.

[15] Appendix S5, Figures S2–S10 show the distribution of treatment effects on these explanatory variables across Muslims and non-Muslims for each event.

Table 2: Effect of intergroup interaction on *GCS* estimated using T-learner using Lasso with 10-fold CV. We separately report average treatment effects for each group and report bootstrapped standard errors in parentheses.

Event	#Users	$\hat{M}_0$		$\hat{M}_1$		$mean(\Delta GCS)$		Treatment Effect			
		R^2	MSE	R^2	MSE	Control	Treated	$\hat{\tau}_{All}$	$\hat{\tau}_{Muslim}$	$\hat{\tau}_{non-Muslim}$	$\Delta\hat{\tau} = \hat{\tau}_{Muslim} - \hat{\tau}_{non-Muslim}$
Janata Curfew	4671	0.091	0.892	0.013	1.088	0.003	-0.027	-0.055	-0.240	-0.038	-0.204
								(0.003)	(0.009)	(0.003)	(0.010)
Tablighi	6946	0.344	0.642	0.345	0.720	-0.001	0.006	-0.160	0.044	-0.181	0.225
								(0.002)	(0.006)	(0.002)	(0.007)
Migrant Deaths	6387	0.116	0.855	0.068	1.081	0.002	-0.013	-0.093	-0.099	-0.093	-0.006
								(0.002)	(0.008)	(0.002)	(.008)
Coronil Launch	4622	0.240	0.753	0.099	0.931	0.007	-0.035	-0.050	-0.082	-0.047	-0.034
								(0.003)	(0.009)	(0.003)	(0.010)
Exam Satyagraha	3497	0.185	0.781	0.119	0.997	0.012	-0.053	-0.053	-0.026	-0.056	0.030
								(0.003)	(0.013)	(0.002)	(0.013)
GDP Contraction	3792	0.146	0.852	0.042	0.935	-0.012	0.051	0.051	-0.010	0.056	-0.066
								(0.004)	(0.012)	(0.004)	(0.010)
Bihar Manifesto	1989	0.066	0.930	-0.025	0.985	0.007	-0.028	-0.086	-0.180	-0.080	-0.100
								(0.006)	(0.033)	(0.006)	(0.032)

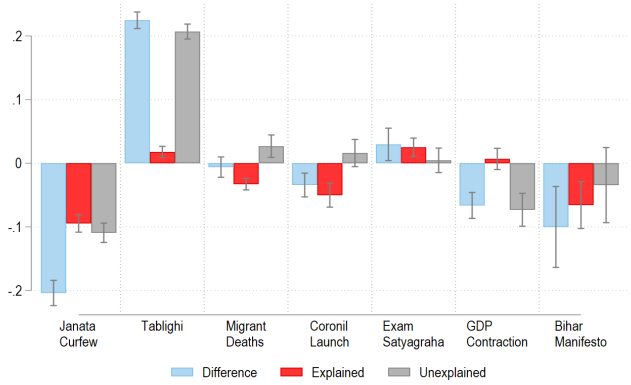


Figure 2: Decomposition of difference in the effect of interaction on *GCS* between Muslims and non-Muslims using Oaxaca-Blinder decomposition. The red bars show the extent to which the effect is explained by topics and emotions. The error bars represent 95% confidence intervals.

Appendix E, Figure 6 decomposes the explained component into contributions of emotions and topics. For Janata Curfew, 81% of $\Delta\bar{\hat{\tau}}$ is explained by valence. This is because, for this event, $\beta_{NM}^{valence}$ is negative—i.e. an increase in valence due to intergroup interaction is associated with a decrease in *GCS*—and the difference $\bar{\hat{\tau}}_{M}^{valence} - \bar{\hat{\tau}}_{NM}^{valence}$ is positive. Additionally, 29% of $\Delta\bar{\hat{\tau}}$ is explained by the topic China & Global for this event. In contrast joy, sadness, and Politics-Religion topic have countervailing effects, i.e. they pull $\Delta\bar{\hat{\tau}}$ towards zero. In case of the Tablighi incident, valence (9%) and joy (14%) explain an important share of $\Delta\bar{\hat{\tau}}$ while anger has a countervailing contribution (-11%). For both the politically salient events—Exam Satyagraha and Bihar Manifesto—the differential effects on Politics-Religion and Socio-Economic topics explain $\Delta\bar{\hat{\tau}}$.

5 CONCLUSION

Our study explores the complex relation between intergroup interactions and polarization between religious groups on social media in light of events during the COVID-19 pandemic. We investigate whether these interactions serve as bridges that mitigate polarization or barriers that exacerbate it. We use a novel measure of group conformity based on contextualized embeddings to uncover a compelling narrative. Consistent with our hypotheses, intergroup interactions generally reduce polarization, though this effect is less pronounced for individuals with stronger group conformity (high GCS). This might be because users holding more extreme positions might be less receptive to outgroup perspectives. Further, in the case of communal events, inter-group animosity may lead to an adverse effect of interactions. We find consistent results—intergroup interactions increase the group conformity for the minority Muslim individuals during the communal Tablighi event. Finally, in the context of political events such as Exam Satyagraha and Bihar Manifesto, intergroup interactions amplify the polarization of politically inclined individuals. Additionally, we leverage a well-known decomposition method to explain the differences in average treatment effects of interaction on group conformity across the two religious groups in terms of effects on emotions and topics of discussion.

More generally, our work highlights the importance of context-aware metrics and nuanced approaches to studying polarization dynamics and its real-time monitoring. This can help inform policies to mitigate increases in polarization and foster healthier social media ecosystems. For instance, by incorporating this metric into real-time recommendation algorithms, these platforms could promote cross-pollination between demographic groups—especially during non-communal events. Importantly, our framework has broad applicability beyond the Indian context and religious polarization. For example, it can be used to estimate speech partisanship in the US Congress or polarization on other social media platforms where some measure of group identity (e.g., demographic or ideological) is known or can be inferred. In line with previous studies that utilize tweet content for predicting group identity [42], the GCS score can also help improve existing name classification algorithms.

ACKNOWLEDGMENTS

We thank the anonymous referees for valuable comments and helpful suggestions. This work is supported in part by NSF under grant No. 2217023, and DARPA under contract number HR001121C0168.

SUPPLEMENTARY MATERIAL

We provide replication code at https://doi.org/10.7910/DVN/NE3DJL and supplementary appendices at https://arxiv.org/abs/2402.11895.

REFERENCES

[1] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review* 110, 3 (2020), 629–676.

[2] Feyaad Allie. 2023. Facial Recognition Technology and Voter Turnout. *The Journal of Politics* 85, 1 (2023), 328–333.

[3] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice. (1954).

[4] Nejla Asimovic, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. 2021. Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences* 118, 25 (2021), e2022819118.

[5] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[6] Ricardo Barandela, José Salvador Sánchez, Vicente Garcıa, and Edgar Rangel. 2003. Strategies for learning in class imbalance problems. *Pattern Recognition* 36, 3 (2003), 849–851.

[7] Michal Bauer, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts. 2016. Can war foster cooperation? *Journal of Economic Perspectives* 30, 3 (2016), 249–274.

[8] Samuel Bazzi, Arya Gaduh, Alexander D Rothenberg, and Maisy Wong. 2019. Unity in diversity? How intergroup contact can foster nation building. *American Economic Review* 109, 11 (2019), 3978–4025.

[9] Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*. 48–53.

[10] Alan S Blinder. 1973. Wage discrimination: reduced form and structural estimates. *Journal of Human resources* (1973), 436–455.

[11] Francesco Bonchi, Edoardo Galimberti, Aristides Gionis, Bruno Ordozgoiti, and Giancarlo Ruffo. 2019. Discovering polarized communities in signed networks. In *Proceedings of the 28th acm international conference on information and knowledge management*. 961–970.

[12] Braz Camargo, Ralph Stinebrickner, and Todd Stinebrickner. 2010. Interracial friendships in college. *Journal of Labor Economics* 28, 4 (2010), 861–892.

[13] Rochana Chaturvedi and Sugat Chaturvedi. 2024. It's All in the Name: A Character-Based Approach to Infer Religion. *Political Analysis* 32, 1 (2024), 34–49.

[14] Sugat Chaturvedi, Sabyasachi Das, and Kanika Mahajan. 2024. When Do Gender Quotas Change Policy? Evidence from Household Toilet Provision in India. *Economic Development and Cultural Change* (2024). https://doi.org/10.1086/729342

[15] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, Vol. 5. 89–96.

[16] Sabyasachi Das. 2023. Democratic backsliding in the world's largest democracy. *Available at SSRN 4512936* (2023).

[17] Saloni Dash, Dibyendu Mishra, Gazal Shekhawat, and Joyojeet Pal. 2022. Divided we rule: Influencer polarization on Twitter during political crises in India. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 135–146.

[18] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2970–3005. https://doi.org/10.18653/v1/N19-1304

[19] Xiaohan Ding, Michael Horning, and Eugenia H Rho. 2023. Same Words, Different Meanings: Semantic Polarization in Broadcast Media Language Forecasts Polarity in Online Public Discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 161–172.

[20] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2023. Non-Polar Opposites: Analyzing the Relationship Between Echo Chambers and Hostile Intergroup Interactions on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 197–208.

[21] Georgy Egorov, Ruben Enikolopov, Alexey Makarin, and Maria Petrova. 2021. Divided we stay home: Social distancing and ethnic diversity. *Journal of Public Economics* 194 (2021), 104328.

[22] Joan Esteban and Gerald Schneider. 2008. Polarization and conflict: Theoretical and empirical issues. , 131–141 pages.

[23] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. TwiBot-22: Towards graph-based Twitter bot detection. *Advances in Neural Information Processing Systems* 35 (2022), 35254–35269.

[24] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4485–4494.

[25] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[26] Venkata Rama Kiran Garimella and Ingmar Weber. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and social media*, Vol. 11. 528–531.

[27] Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87, 4 (2019), 1307–1340.

[28] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[29] Pedro Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. 2013. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 215–224.

[30] Raj Kumar Gupta, Ajay Vishwanath, and Yinping Yang. 2020. Global Reactions to COVID-19 on Twitter: A Labelled Dataset with Latent Topic, Sentiment and Emotion Attributes. *arXiv preprint arXiv:2007.06954* (2020).

[31] Yosh Halberstam and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* 143 (2016), 73–88.

[32] Thomas F Homer-Dixon. 2010. *Environment, scarcity, and violence*. Princeton University Press.

[33] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.

[34] Matthew S Levendusky and Dominik A Stecula. 2021. *We need to talk: how cross-party dialogue reduces affective polarization*. Cambridge University Press.

[35] Matt Lowe. 2021. Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review* 111, 6 (2021), 1807–1844.

[36] Rajeshwari Majumdar. 2023. *Reducing Prejudice and Support for Religious Nationalism Through Conversations on WhatsApp*. Technical Report.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[38] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25, 3 (2015).

[39] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. 2018. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*. 369–378.

[40] Ronald Oaxaca. 1973. Male-female wage differentials in urban labor markets. *International economic review* (1973), 693–709.

[41] Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American journal of political science* (1985), 357–384.

[42] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*. 1534–1545.

[43] Gautam Rao. 2019. Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools. *American Economic Review* 109, 3 (2019), 774–809.

[44] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

[45] Tamar Saguy, Nicole Tausch, John F Dovidio, and Felicia Pratto. 2009. The irony of harmony: Intergroup contact can produce false expectations for equality. *Psychological science* 20, 1 (2009), 114–121.

[46] Erik Santoro and David E Broockman. 2022. The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science advances* 8, 25 (2022), eabn5515.

[47] Linda R Tropp and Thomas F Pettigrew. 2005. Relationships between intergroup contact and prejudice among minority and majority status groups. *Psychological Science* 16, 12 (2005), 951–957.

[48] William J Youden. 1950. Index for rating diagnostic tests. *Cancer* 3, 1 (1950), 32–35.

A USER NAME CLEANING AND FILTERING ORGANIZATIONS

We transliterate Twitter usernames from Indic languages Hindi, Bengali, Gujarati, Punjabi, Malayalam, Kannada, Tamil, Telugu, Oriya, Marathi, Assamese, Konkani, Bodo, Nepali, and Urdu to English using Indic-trans tool [9].[16] To drop non-personal names, we construct a name part dictionary using names from a 3% random sample of eligible voters from Indian electoral rolls and the Rural Economic & Demographic Survey (REDS) data collected by the National Council of Applied Economic Research. We use the 3% sample due to data availability constraints. This is a large sample comprising over 25 million voter names and their parent's/spouse's names (15,431,765 unique names) out of over 800 million total voters. For every Twitter user, only name parts that occur in the constructed name part dictionary are retained. We further manually scan names of users who either have more than 20,000 followers, tweeted more than 60 times, or whose names contain any of the organization-related keywords provided in Table 3 and drop those having non-personal names from this list.

Table 3: Keywords used to filter out organization names from the tweeters after lower-casing the usernames

group, team, organization, foundation, official, college, university, universities, fan, fc, school, institute, institutions, chamber, brand, service, board, bureau, gov, division, technology, consult, khabar, voice, collector, medical, health, mirror, journal, chronicle, post, daily, times, today, channel, temple, station, bjp, congress, council, business, shop, party, bollywood, cinema, academy, center, centre, state, collective, association, indian, group, sangh, NGO, RBI, online, cooperative, retail, .com, .in, .edu, .org, hospital, research, solution, department, bank, adani, fan, HSBC, sena, dpro, logic, tech, district, state, work, CPI, INC, BSP, AAP, CPM, NCP, BJP, trust, govt, Prakashan, corporation, socialist, communist, committee, janta

B MUSLIM CLASSIFICATION THRESHOLD

To choose the Muslim classification threshold and to assess the performance of our models, we first manually annotate a sample of thousand names as Muslim or non-Muslim. We select this subset by splitting all the names into equal-width bins after sorting them based on the Muslim score. We use 20 bins of width 0.1 each. The first bin includes names with scores below -0.9 and the last with 0.9 and above. We then randomly sample 50 names from each bin and annotate their perceived religion. Thereafter, we analyze the points that maximize the geometric mean (G-mean) ($\sqrt{sensitivity * specificity}$) [6] and the Youden's J-index ($sensitivity + specificity - 1$) [48]. Both these measures are used to determine the optimal cut-points that maximize the predictive performance of each class while keeping it balanced. We choose the threshold of 0.3 as the decision boundary based on these statistics (see Figure 3).

[16]Available at https://github.com/libindic/indic-trans.

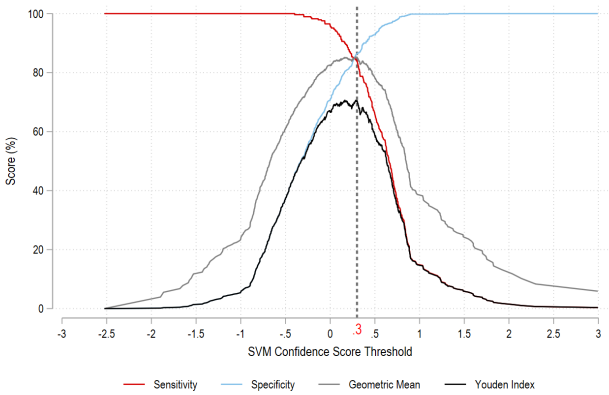
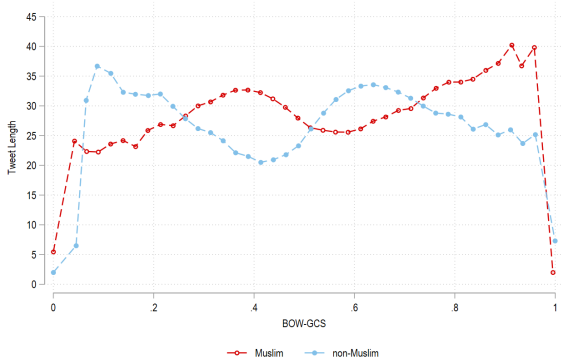


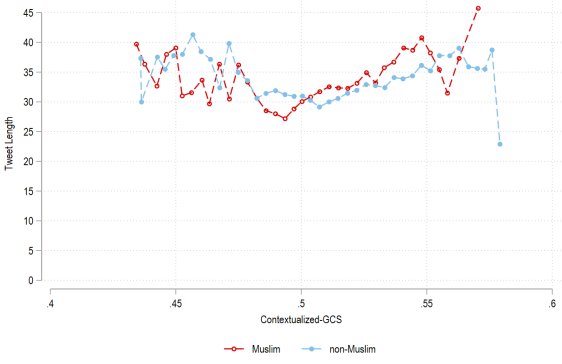
Figure 3: Sensitivity, specificity, Youden index, and geometric mean by prediction threshold.

That is, if the Muslim score is greater than 0.3, we classify a name as Muslim, and non-Muslim otherwise.[17]

C TWEET LENGTH: BOW-GCS AND GCS



(a) BOW-GCS



(b) Contextualized-GCS

Figure 4: Tweet length and GCS.

[17]We also find a very common non-Muslim name Abhishek classified as Muslim, and manually classify this as non-Muslim. We also experiment with a threshold of zero and find qualitatively similar results.

D SUMMARY STATISTICS

Table 4: Descriptive statistics for the dataset. For each event-specific subset, 30-day pre-event averages are reported for the covariates along with standard deviations in parentheses. *M: Muslim, NM: Non Muslim

	EVENT	Janata Curfew	Tablighi	Migrant Deaths	Coronil Launch	Exam Satyagraha	GDP Contraction	Bihar Manifesto
	Date	Mar 22 2020	Mar 31 2020	May 8 2020	Jun 23 2020	Aug 23 2020	Aug 31 2020	Oct 22 2020
	Muslim %	8.39	9.11	7.86	7.81	8.38	8.07	6.54
Interact	Overall%	10.30	12.14	15.06	16.96	18.84	19.09	20.21
	Muslim %	54.34	53.08	62.55	67.04	65.87	69.61	75.38
	Non Muslim %	6.26	8.03	11.01	12.72	14.54	14.66	16.35
GCS	Overall	0.5007	0.5008	0.5014	0.5012	0.5007	0.5007	0.5009
		(0.002)	(0.0027)	(0.0026)	(0.0023)	(0.0024)	(0.0022)	(0.0024)
	*M Interact	0.4997	0.5003	0.5005	0.4996	0.5002	0.4999	0.4994
		(0.0021)	(0.0028)	(0.0035)	(0.0025)	(0.0031)	(0.0025)	(0.0022)
	*NM Interact	0.5006	0.5005	0.5008	0.5016	0.5009	0.5009	0.5009
		(0.0022)	(0.0027)	(0.0027)	(0.0021)	(0.0021)	(0.0021)	(0.0020)
	*M non-Interact	0.5007	0.5019	0.5018	0.5010	0.5011	0.5008	0.5009
		(0.0022)	(0.0034)	(0.0033)	(0.0028)	(0.0033)	(0.0030)	(0.0024)
	*NM non-Interact	0.5007	0.5007	0.5015	0.5012	0.5007	0.5008	0.5009
		(0.0020)	(0.0026)	(0.0025)	(0.0023)	(0.0023)	(0.0022)	(0.0024)
Topics	COVID Response	0.69	0.64	0.59	0.58	0.59	0.57	0.59
		(0.27)	(0.22)	(0.19)	(0.19)	(0.2)	(0.19)	(0.19)
	Politics-Religion	0.14	0.16	0.21	0.19	0.22	0.24	0.2
		(0.22)	(0.2)	(0.22)	(0.23)	(0.27)	(0.29)	(0.28)
	China & Global	0.06	0.07	0.04	0.07	0.02	0.02	0.03
		(0.28)	(0.22)	(0.19)	(0.16)	(0.18)	(0.18)	(0.16)
	Socio-Economic	0.11	0.13	0.17	0.15	0.16	0.16	0.17
		(0.16)	(0.17)	(0.2)	(0.23)	(0.25)	(0.25)	(0.24)
Emotions	Valence	0.45	0.46	0.47	0.46	0.46	0.46	0.47
		(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)
	Fear	0.45	0.45	0.44	0.45	0.45	0.45	0.44
		(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
	Sadness	0.41	0.42	0.41	0.41	0.42	0.42	0.41
		(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
	Joy	0.3	0.3	0.31	0.31	0.3	0.3	0.31
		(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)
	Anger	0.44	0.44	0.44	0.44	0.44	0.44	0.43
		(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Ego-Net	Followers	2496.63	2401.09	2867.94	3315.16	2954.83	2766.27	3823.92
		(8586.73)	(8384.17)	(9929.8)	(12296.33)	(10296.9)	(9375.01)	(11418.67)
	Friends	936.18	935.06	988.95	1004.55	986.22	966.25	1137.37
		(1417.11)	(1439.22)	(1653.84)	(1532.91)	(1519.08)	(1431.95)	(1766.76)
Engagement	Retweets	1.93	2.27	2.41	2.57	3.51	4.28	2.87
		(19.35)	(16.01)	(16.4)	(16.28)	(22.88)	(38.91)	(15.4)
	Fraction of replies	0.66	0.61	0.53	0.51	0.51	0.54	0.48
		(0.31)	(0.31)	(0.35)	(0.37)	(0.38)	(0.39)	(0.39)
	Tweet frequency	7.5	12	16.07	13.08	10.73	9.89	9.34
		(10.08)	(16.93)	(25.58)	(24.94)	(19.72)	(19.96)	(17.6)
	Account days	2493.81	2476.94	2490.63	2539.8	2465.07	2524.91	2635.2
		(1253.79)	(1244.03)	(1266.25)	(1292.43)	(1353.38)	(1335.86)	(1335.84)
	Tweeters	4671	6946	6387	4622	3497	3792	1989

E TREATMENT EFFECT HETEROGENEITY

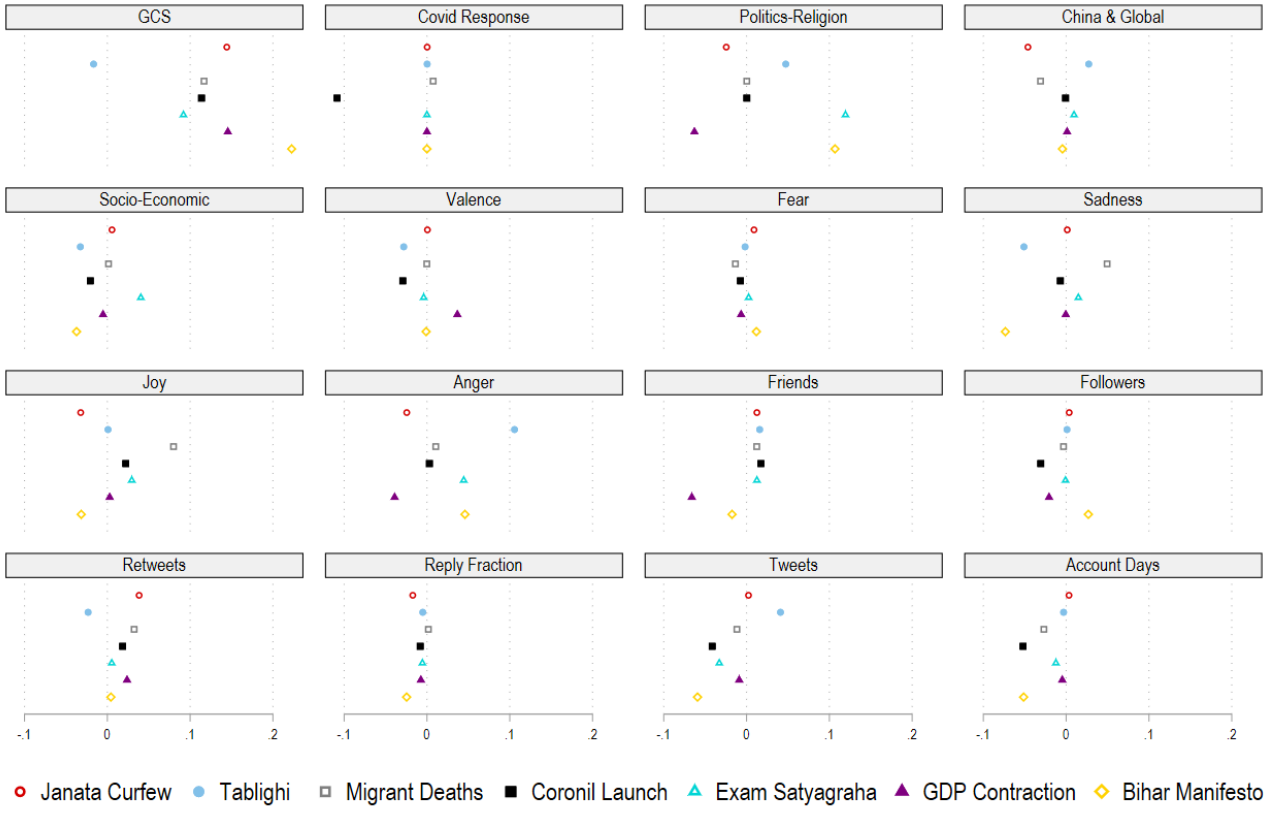


Figure 5: Coefficient plot of covariates when treatment effect is regressed on them.

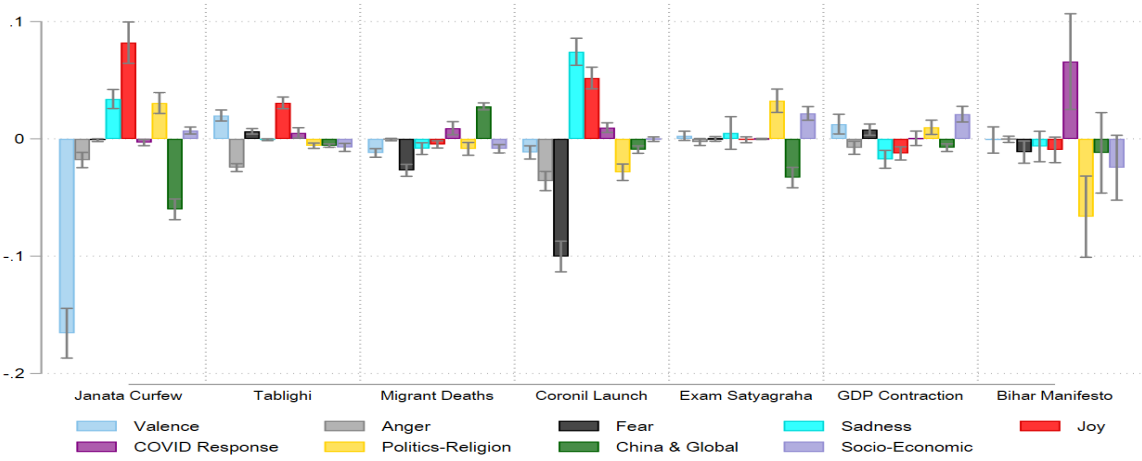


Figure 6: Contribution of emotions and topics towards the explained component of difference in the effect of interaction on ΔGCS across Muslims and non-Muslims using Oaxaca-Blinder decomposition. Errors bars show 95% confidence intervals.