Political mud slandering and power dynamics during Indian assembly elections

Sarah Masud[1] and Tanmoy Charaborty[2]

[1]Department of Computer Science & Engineering, Indraprastha Institute of Information Technology Delhi, New Delhi, India, 110020
[2]Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India, 110016

July 18, 2023

Abstract

Political parties often engage in verbal swordplay, which worsens during elections. Free and fair elections are the pillar of a democratic society. Come election season, all media outlets, including the Internet, are buzzing with politically-charged content. We curated $46k$ posts from Twitter between January-March, 2022 to examine political content during the Indian assembly elections of February 2022. We begin our analysis by manually labeling $1.7k$ posts for different forms of attack and employ the annotation to examine political attacks against defamatory hashtags and name-calling. It anecdotally lends itself to the target's online reputation. Similarly, we quantify the power dynamics of self-promotion and negation pinned on the ruling party before and after elections. To aid large-scale analysis, we obtain pseudo-labels for the rest of the dataset via training a political attack detector for the Indic setting. Subsequently, we observe that the patterns detected via manual annotations hold at scale too. Our analyses and findings aim to educate the citizenry about the quality of political discourse on Indian Twitter.

Political Attack, Indian Twitter, Assembly Elections

1 Introduction

In the contemporary world, apart from the conventional means of electioneering, social media has emerged as a cheap and fast medium of information dissipation for political parties (1, 2). However, unregulated political interactions on social media (3) can extend to trolling, polarization (4), hate speech (5, 6), and even violence (7). We take this opportunity to analyze the quality and quantity of political discourse on Twitter among other platforms (8) given its rising popularity in India. Given data collection constraints, it is difficult to comment on general trends on how the political discourse shapes social media. However, under the lens of assembly elections, our study aims to examine correlational evidence to various online and offline events.

India is the largest democracy and fifth largest economy in the world. Political stability in India acts as a canary in a coalmine regarding global socio-political cohesion. The current study examines recent Indian assembly elections, February 2022, held in five states of India – Uttar Pradesh (UP), Punjab, Goa, Uttarakhand, and Manipur. Before the elections, Bhartiya Janta Party (BJP) was the majority in Goa, UP, Uttarakhand, and alliance-based power in Manipur. They came into power in four states, winning a majority in Goa, Manipur, and Uttarakhand, and retained power in UP under an alliance. Meanwhile, in Punjab, the power shifted from Indian National Congress (INC) to Aam Aadmi Party (AAP). With the general elections scheduled for 2024, the assembly elections 2022 became a litmus test for political parties' future strategies (9).

Our analyses of Twitter posts (interchangeably called tweets) extend upon the established norms of Indian political discourse (10, 11). Given that it is difficult to predict election results, and the data we collected is only a subset of the overall electoral social media data, we analyze the patterns anecdotally. We begin by studying the general trends of velocity and veracity of engagements and then move towards a granular analysis of different types of attacks and self-promotion (2). A political attack is treated as a sub-category of offense, non-overlapping with identity-based attacks such as hate speech (5, 12, 13, 14). We associate the observed online patterns with real-world events to establish the power dynamics between the ruling and opposing groups. This study aims to showcase how elections shape the political discourse on social media.

2 Volume of Attack

For reader's reference, the list of political parties analyzed in this study is provided in Section 6

Labeling attacks. Analysis of political attacks requires human annotation. We achieve the same via manual labeling of $1.7k$ posts and term this subset as *annotated samples* (Figure 1A). We then pseudo-label the rest of the datasets for large-scale analysis via a classification model. This group of samples is termed *predicted samples* (Figure 1D). Details about annotation and modeling can be found in Materials and Methods and *SI Appendix* (Sections 1 and 2).

Frequency of Attack. Since BJP and INC are India's two largest political parties, they were also the most active on Twitter, making up for 75% of the tweets produced in our dataset. BJP and INC contributed to $23,879$ and $11,783$ tweets, respectively. Meanwhile, Bhagwa Kranti Sena (BKS) and Azad Samaj Party (ASP) displayed an intriguing social media presence. While BKS and ASP contributed a low volume of tweets, their posts still received considerable retweets and likes. Upon analyzing some of the content of ASP and BKS from

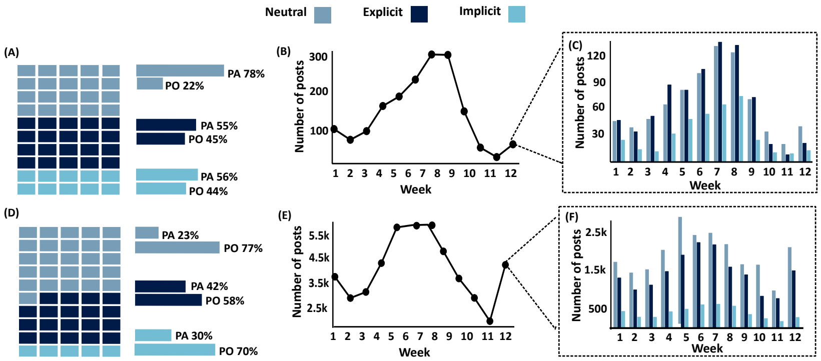


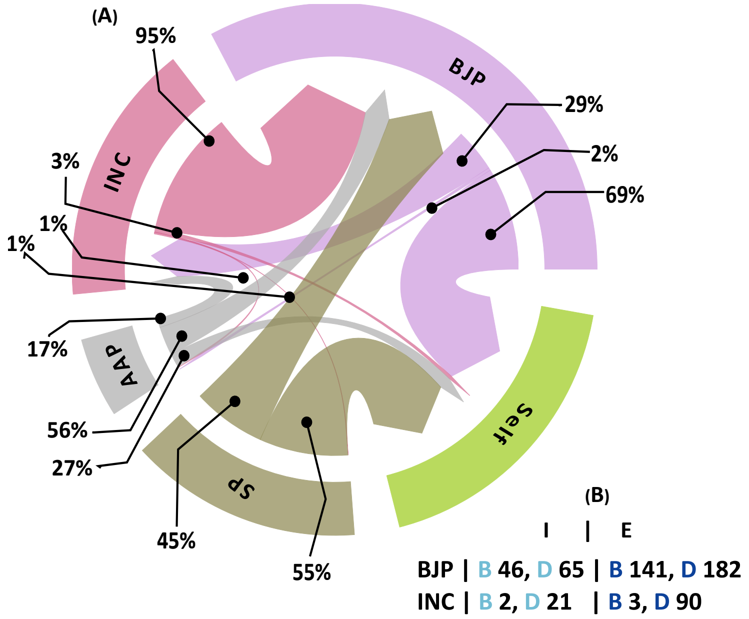
Figure 1: Overview of the political attacks from January-March 2022 curated from Indian Twitter, with a breakdown of neutral, explicit, and implicit remarks. The top row highlights the pattern of attacks for the manually-annotated samples; the bottom row highlights the same set of patterns for predicted samples. The curated dataset has 695 (*resp.* 23, 838) neutral, 696 (*resp.* 17, 771) explicit, and 329 (*resp.* 4, 858) implicit instances of manually annotated (*resp.* model predicted) samples of political attacks represented in pictogram A (*resp.* D). Each box presents 1% of the respective class, totaling $1.7k$ manual and $45.3k$ predicted samples, respectively. The two user groups in the curated dataset are Political Party handles (PY) and Politicians (PO). As the elections were held in phases throughout February 2022, we also study the week-wise breakdown of attacks where the x-axis represents the nth week, and the y-axis represents the volume of posts. (B) and (E) capture the week-wise volume of posts from January-March based on our annotated and predicted samples, respectively. Meanwhile, (C) and (F) represent a granular weekly proportion of neutral, explicit, and implicit attacks obtained from our annotated and predicted samples.

our annotated samples, we observe that BKS either posts tweets in support of the BJP or directly attacks other parties. Meanwhile, tweets by ASP are all politically attacking BJP. The percentage of attacks vs. neutral within full datasets in BKS (48.3%) and ASP (60%) further corroborates this pattern. An example of BKS supporting BJP in a neutral manner which received 3, 371 and 1, 911 retweets and likes is the post "पूरा यूपी डोल रहा है, योगी योगी बोल रहा है! " (Entire UP is mesmerized and is enchanting Yogi Yogi!). A similarly high interaction post explicitly attacking BJP coming from ASP is "*General Dyer in the guise of a sage crossed all limits of cruelty. So ruthlessly that even the sticks, the British started to feel embarrassed. First, students were beaten fiercely; now, FIR was lodged against 1000 students. The youth will answer Yogi Ji, who is preparing to criminalize the country's future. #RRBNTPC_Scam*". It garnered 4, 663 (15, 156) in retweets (likes).

Proportion of attacks. Among the annotated samples (Figure 1A), the ratio of attack to neutral speech is 3 : 2. It is anticipated as the data was curated during the election season. The week-wise spread of annotated attacks peaked in February when the actual elections were held and dwindled post that. This trend is visible both at the overall volume level (Figure 1B) and label-breakdown (Figure 1C). Interestingly, trends from the annotated samples translate to the predicted samples obtained via pseudo-labeling (see Section 6). Among predicted samples, we observe a ratio of neutral vs. attacking being $1:1$. The spread of attacks per week (Figure 1E) and per label (Figure 1F) again highlights February as the most active month. As expected in both manual and predicted samples, the proportion of attacks dwindles, or neutral posts increase once elections end. Direct attacks in manual and predicted samples overshadow implicit ones by $2:1$ and $3:1$, respectively. Interestingly, our analysis finds that the mean retweets/likes in explicit posts are significantly more (t-test) than in implicit posts for pseudo-labeled samples but not significant for manually annotated samples. For pseudo labels, the t-test for median retweets/likes yields p-value of $1.030e^{-5}/1.164e^{-17}$ and an effect size of 0.654/1.366. Meanwhile, for manually annotated samples, the t-test for median retweets/likes yields p-value of 0.79/0.51 and effect size of 0.11/0.28. We pin this difference in a pattern on the pseudo-labeled set's $\approx 20X$ larger size. When we randomly selected 1.7k samples from the pseudo labeled set and ran the t-test, the best and effect size values for retweets/like mimic that of the manually annotated samples with p-value of 0.684/0.006 and effect size of 0.176/1.301. It points toward the fact that, in general, explicit posts are more likely to receive more engagement than their implicit counterparts, as the former is easier to comprehend by a larger audience. These findings are in line with existing literature on the spread of hateful content (15).

3 Advertising or attacking

Given the upper hand that BJP yields at a central and state level, we observe that their hashtags for self-promotion heavily outnumber those for the demotion of the opposition. While these hashtag frequencies are representative of the curated 46k post, the proportions should also scale in the wild. They employ self-promoting hashtags like #यूपीमांगेभाजपा (UPDemandsBJP) and #TripuraWelcomesModiJi with frequencies of 120 and 35. While other hashtags like #भगवामय_उत्तराखंड (Uttrakhand will be painted orange) and #चप्पा_चप्पा_भाजपा (BJP at every nook and corner) with frequencies of 39 and 75 have an undertone of intimidation; they still display dominance and self-promotion. Meanwhile, they rarely utilize explicitly-targeted hashtags such as #अपराधी_वाली_सपा (SP's inherit Terrorism) or #KejriwalAgainstHindus with frequencies of 5 and 3, respectively. Other parties like AAP, INC, and SP balance their resources for attacking the ruling party and self-promotion. The hashtags they employ reflect the same. While attacking BJP, they adopt a combination of explicit and implicit hashtags such as #BJPKeGunde (BJP's goons), #DalitVirodhiModi (Anti Dalit Modi), and #संविधानद्रोही _BJP (BJPisAntiConstitution) with frequencies of 3, 36 and 1, respectively. Note we employ the presence of "#" at the start of a word to capture unique hashtags and their frequencies.



	I	E
BJP	B 46, D 65	B 141, D 182
INC	B 2, D 21	B 3, D 90

Figure 2: An overview of promotion and demotion by four parties – BJP, INC, AAP, and SP. (A) The weighted and directed chord diagram represents the dynamics of self-promotion and denunciation/attack on opponents among the four parties. The interactions are based on the volume of manually-annotated promotions and demotions (*aka* advertising vs. attacking) of 1.7*k* tweets. Given that parties under consideration are not allies, we can safely assume that comments directed at other parties are not promotional. Self-promotion is captured by edges directed at the self (green) chord. We also annotate each chord with the percentage breakdown of tweets involved in a particular action, e.g., $X \xrightarrow{p\%} Y$ represents the p% of the tweets by X is directed towards Y. (B) A breakdown of denunciation by BJP and INC. Here I and E represent the implicit and explicit attack labels; meanwhile, B and D capture whether the attack is either promotion and demotion or purely denouncing.

To further analyze the distribution of attacks and promotions among 1.7*k* manual annotations, we add another set of labels and mark the entities involved in promotion and demotion (see Section 6). We mark self-promotion for AAP, INC, SP, and BJP, denouncing others, both or none. With INC, we observe a ratio of 7 : 1 for attacking vs. advertising. INC is currently the largest opposition party. Therefore, INC likely criticizes the ruling BJP at national and regional levels. This behavior is corroborated via Figure 2, in which we observe a higher density of attacks directed from INC to BJP. From 2017 to 2022, INC's vote share went down from 28.7% to 23.7% in Goa, 38.8% to 23.1% in Punjab, 6.3% to 2.4% in UP, and 35.3% to 16.9% in Uttarakhand. Their only success came from Uttarakhand, where their share increased from 33.8%

to 38.2%. Despite being one of India's largest and oldest parties, INC has recently seen a decline in voter share and the political power they yield. It could be one of the reasons why BJP responds with a single attack for every three attacks initiated by INC (Figure 2).

Note on the BJP-INC Dynamics. The observation that BJP attacks INC lesser in proportion at first seems counterintuitive given that BJP and INC attack each other regularly and vigorously. However, one must interpret this observation exclusively under the elections with a focus on UP. In UP's last assembly elections (2017), INC won only 7 seats while BJP won 312. Similar trends were observed in Goa, Manipur, and Uttarakhand. BJP was further able to hold its popularity until the next elections. It means that in 2022, they contested elections unassailably, represented by the 69% of their posts being about rallies, visits, or self-promotional information. In the rest of their denunciation of other parties, they still spend the majority on INC (29%). This proportion of attacks becomes more prominent when we look only at BJP and INC's content concerning manually annotated attack labels. In absolute value, BJP still contributes significantly (compared to INC) to implicit (65 vs. 21) and explicit denunciations (182 vs. 90) (Figure 2*B*). To have any significant impact on the ground, INC plays defense, leading to an increase in its attack against BJP. While BJP continues to produce many attacks, as known anecdotally, during the election, they focus more on their popularity and publicity. They can denounce from a secure position leading to a skew in the proportion of attacks by these parties, which could be one reason for the observed behavior.

4 Boon or bane of name-calling

Political leaders are often given nicknames depending on whether they receive the name from their supporters or opposition. For example, Prime Minister Narendra Modi is affectionately referred to as "Sher" (*Lion*) by his supporters; meanwhile, INC leader Rahul Gandi is frequently called "Pappu" (*dim-witted*) by the opposition. Back in the year 2017, when opposition leader Rahul Gandhi remarked, "*They say, you give me potatoes, and I will pass them through a machine turning them into gold*," he was rhetorically referring to prime minister Narendra Modi. However, a cropped version of Rahul's speech where only the words "*you give me potatoes, and I will pass them through a machine turning them into gold*" was widely circulated on the web. A political gibe backfired for Rahul Gandhi, who, to date, is trolled for this clip. Such is the nature of Indian politics! Similarly, during the Assembly elections, we notice some new and old nicknames surface on Twitter.

Yogi's bulldozer image. BJP leader and UP's CM Yogi Adityanath is often referred to as "Baba" (sage), given that he was a saint before joining politics. In our study, we observe the term "baba" being used as a source of affection as well as irony. Consequently, "Bulldozer Baba" (the bulldozer saint) refers to his act of bulldozing the properties of apparent criminals and unlawful citizens. While the opposition was asking if bulldozers and the destruction of the property could improve the financial conditions in UP, his supporters were presenting him with gold platted bulldozers to cement his rule with an iron fist. Upon Adityanath's re-election as the CM of UP in 2022, his supporters flooded social media with images and videos of bulldozers. Some were even seen standing atop bulldozers and chanting victory to symbolize that nothing could come in the way of the bulldozer, like the determination of Yogi Adityanath. The tweet, "बुलडोज़र बाबा से मन का सवाल। क्या बुलडोज़र से बेरोज़गारी दूर हो जायेगी? किसान को फसल का दाम मिल जायेगा? महँगाई ख़त्म हो जायेगी? " (I have the following question in my mind for Bulldozer Baba, will bulldozers solve unemployment, give farmers their due and reduce inflation?), captures the direct attack employing the nickname "Bulldozer Baba". On the other hand, we also see examples where tweets put a positive spin on the nickname (a) "बुलडोजर का बढ़ता क्रेज। गोरखपुर के प्रमुख व्यापारी ने उत्तर प्रदेश मुख्यमंत्री योगी आदित्यनाथ जी को चांदी का बुलडोज़र भेंट किया। " (Bulldozer is gaining popularity. A prominent businessman from Gorakhpur gifts Yogi Adityanath a silver-plated bulldozer), and (b) "यूपी में सरकार बनाने जा रही बीजेपी, हेमा मालिनी बोलीं- बुलडोज़र के आगे कुछ नहीं आ सकता " (*BJP will be forming the government in UP, Hema Malini says - Nothing can stand in the bulldozer's way.*)

The tussle in Punjab. While INC aimed to retain its majority in Punjab, BJP, and AAP aimed to overthrow the incumbent. As such, the three groups indulged in hefty name-calling. Opposition extensively utilized the term "RetaChorChanni" (Sand Mafia Channi) to refer to Charanjit Singh Channi, Punjab's then CM, for his apparent implication in the case of sand mining scams and corruption. The damaging #RetaChorChanni campaign against Channi proved disadvantageous. What started as a slandering remark got bolstered when reports of corruption probes against him and his associates came to light. Some posts attacking his leadership were expressed as (a) "जिस चन्नी को राहुल गांधी कोहिनूर बताते है, पंजाब की जनता उसे रेता चोर बुलाती है। सोचने वाली बात है। सिर्फ 111 दिन के राज में घर से 10 Crore Cash बरामद हुए #RetaChorChanni" (The very Channi whom Rahul Gandi calls a gem, the people of Punjab call him a Sand Mafia. In his 111 days of ruling, it is quite astonishing how he amassed 10 Crore Cash.) and (b) "CONgRSS promotes corrupt people. Their top leadership facilitates corrupt and incompetent people. Siddhu compromised on a Sand Mafia being a CM face from your Congress. #RetaChorChanni #RetaChorMachayeShor" (Sand Mafia crying wolf). The opposition's constant remarks on #RetaChorChanni even led news media to question his importance in the INC-alliance in Punjab with headlines like "पंजाब में कांग्रेस के लिए चन्नी जरूरी या मजबूरी? " (Is Chennai an asset or a burden for Congress in Punjab?).

In another series of events-related events, various opposition parties started calling AAP's leader Arvind Kejriwal a "terrorist." Having established their mantle in Delhi state elections, AAP hoped to extend their party and replicate the results in Punjab and Goa. Puns suggest that the Aam Aadmi Party (the party of/by commoner) is rather "अरविंद आतंकवादी पार्टी (Arvind Terrorist Party), and this was widely circulated. The tweet can be seen as equating Kejriwal and his party to a terrorist organization. He turned this smear campaign to his advantage by saying he is the world's first "sweet terrorist" who works for his people's welfare. He put a positive spin on the smear campaign, showcasing his parties' work in Delhi and what they hope to replicate and improve in Punjab. With banters such as "जिसको पप्पू और गप्पू आतंकवादी कह रहे है, वही केजरीवाल जी देश के बच्चो के लिए हज़ारो स्मार्ट क्लासरूम बना रहे है। फर्क साफ है जो भी अच्छी शिक्षा, फ्री बिजली, अच्छे हॉस्पिटल की बात करेगा उससे भाजपा+कांग्रेस मिलकर लड़ेंगी। " (The pappus and gappus are calling me a terrorist. I am that terrorist who has installed smart classrooms for the kids. The difference is clear, anyone who speaks of quality education, free electricity, and good hospitals, BJP+INC, will unite to fight that person.), Kejriwal can be seen mockingly bashing the opposition for trying to sully any leader like him who speaks of the welfare of the people.

The symbolic "Lal Topi". SP leader Akhilesh Yadav always dons a "Lal Topi" (*red cap*) to symbolize his party's socialist ideology and the struggles of the downtrodden. Conversely, the opposition implicitly uses "Lal Topi" to symbolize corruption, red-tapism, and violence. For example, BJP leaders utilize "Laal Topi" to be a connotation of Akilesh's criminal affiliations, and occasionally make remarks like – लाल टोपी का मतलब 'दंगा', लाल टोपी का मतलब 'हिस्ट्रीशीटर', लाल टोपी का मतलब राह चलते नागरिकों के साथ 'रहजनी', लाल टोपी का मतलब किसानों के खेत से 'ट्यूबेल और पंपसेट का चोरी' हो जाना (Red cap represents riots, the red cap represents hardened criminals, the red cap represents looting and snatching on streets, and the red cap represents stealing of hand pumps and tube wells from farm fields).

5 Discussion

Flexing by BJP. A surprising yet apparent trend is the ruling and opposing power dynamics. BJP is the ruling party at the center in four out of five states in these elections (16). They thus participated in the elections from a position of security and comfort, with 69% of their annotated samples being self-promotion (Figure 2). One observes the same regarding hashtags and name-calling employed by them. Opposition parties with fewer resources need to be more dexterous about promoting themselves. They bank on the incumbent's lack of sympathy toward public issues to launch negative campaigning against the BJP. AAP and SP display this behavior.

A case for AAP in Punjab and Goa. AAP's most successful campaign was in Punjab, where they had a landslide victory winning 92 out of 117 seats. They tactfully curated an equal combination of promotional and denouncing hashtags (17). On the one hand, they banked on the incumbency of the existing parties in tackling issues of unemployment, education, and rampant drug usage in Punjab. They employed hashtags like #PunjabDiUmeedAAP (AAP is Punjab's hope) and #AAPKeGovtSchools (Government Schools by AAP), which instilled hope of AAP bringing a systematic change. On the other hand, the alleged involvement of the state's then chief in scams prompted the use of employed attacking hashtags like #DarrGayaChanni (Channi is afraid) and #RetaChorChanni (Sand Mafia Channi). Interestingly, a similar tactic of promoting the so-called "Delhi Model" did not favor AAP in Goa. For the opposition, along with self-promotion and demotion, it is safe to say that the level of incumbency and public dissatisfaction is equally important to swing voters.

Rising popularity of SP. The number of seats won by SP in UP increased from 47 in the year 2017 to 111 in 2022. SP is also the third leading party in our curated dataset regarding the number of tweets with an average retweet and like counts of 650 and 3365, respectively. While other parties could achieve the same numbers on social media, their user engagements did not translate into vote share. The rising popularity of SP in the online world can be both an antecedent and a consequence of its rising popularity. Coupling this with changing political awareness among the rural population in UP, (18), one can expect the general elections to take an interesting turn.

Online attacks reflect offline invective. In worse scenarios, offline invectives and political attacks can lead to internal/domestic extremist groups acting in ill faith (19). During one of the campaigns, Punjab's then CM Charanjit Singh Channi made a statement– "Do not let UP, Bihar ke bhaiya enter Punjab" (Do not allow men from UP and Bihar to enter Punjab). He later clarified that it was meant for the people of Punjab not to let political leaders like AAP from Delhi enter Punjab politics. However, his clarification came in too late. What was supposed to be an implicit attack on AAP ended up as a barrage of explicit attacks against him and other female politicians who accompanied him (20). This incident alone spiked the volume of overall and explicit posts as captured in week 7 of Figures $1E$ and $1F$.

Any hateful attack focused on identity (like religion, caste, or gender) rather than political ideology should be discouraged (21, 22). Barring that, in a vibrant democracy, political attack and criticism are the norms (23). Despite the rich prevalence of direct attacks, most of them walk the thin line of provocation. During elections, noteworthy events in the physical world get magnified online (24). We observe the same in this study. It will be interesting to see how patterns emerging from this election play out in the General Election of 2024.

6 Materials and Methods

Dataset overview. We curated $46k$ tweets amassed biweekly from Jan 1 – Mar 31, 2022, by focusing on 111 political leaders representing 17 political groups and one set of independent candidates. Our dataset includes political leaders from – Aam Aadmi Party (AAP), All India Majlis-E-Ittehadul Muslimeen (AIMIM), Apna Dal (AD), Azad Samaj Party (ASP), Bhartiya Janta Party (BJP), Bhagwa Kranti Sena (BKS), Bahujan Samaj Party (BSP), Goa Forward Party (GFP), Indian National Congress (INC), Jammu and Kashmir Peoples Democratic Party (PDP), Lok Janshakti Party (LJP), Rashtriya Janta Dal (RJD), Rashtriya Lok Dal (RLD), Samajwadi Party (SP), Sanyukt Samaj Morcha Party (SSM), Shiromani Akali Dal (SAD) and Shiv Sena (SHS). We also examine the official Twitter handles of six political parties – BJP, INC, AAP, SP, BSP, and AIMIM. More details about the dataset can be found in Appendix A.1.

Data annotation. Within the larger purview of offensive content, *political attacks can be defined as defamation and accusatory remarks made to project* *the opposing political groups negatively* (25).

A direct/explicit attack targets political affiliations, ideologies, and policies and may employ political name-calling. Instances of name-calling and offensive slurs (about politics) also form direct attacks. Indirect/implicit attacks are expressed with taunt, sarcasm, irony, etc. and may require cultural/contemporary referencing for comprehension. It should be noted that while these political attacks fall under the more extensive umbrella of derogatory content, these political attacks do not qualify as hate speech because politicians and political parties are not vulnerable entities (8, 26). It qualifies as hate only if an attack is on personal identities (race, religion, caste, etc.). Hence in this discussion, we stick to the term political attack.

We annotated 1.7k samples from our datasets for political attacks based on the above definitions. Data annotation was carried out by three annotators who were well-versed in social computing and Indian politics. They were professional annotators, and both of them were female. The annotators were proficient English and Hindi speakers and used Google Translate to understand content in other regional languages.

In the first phase, they separately annotated samples of 100 posts each, refining the definitions of attack until they reached an inter-annotator agreement of 0.70 (Cohen's κ). Following this, each annotator contributed 750 annotations for 1.7k samples labeled as 'neutral,' 'implicit,' or 'explicit.' Here the three labels 'neutral,' 'implicit,' or 'explicit' captures the coarse grain sentiment. For example, the following statement taunts other political parties for the kind of leaders they pay homage, thereby qualifying as an 'implicit attack' – 'BJP: Savarkar और Hedgewar की तस्वीर क्यों नहीं लगाई? Congress: Indira Gandhi, Rajeev Gandhi, Sonia Gandhi की तस्वीर क्यों नहीं लगाई? मैं कहता हूँ कि इन सबकी तस्वीर आप लोग लगा लो, हम तो बाबा साहेबAmbedkar और Bhagat Singh जी की ही तस्वीर लगाएंगे।'

In the second phase, for tweets posted by BJP, INC, AAP, and SP (politicians and party handles), the annotators also marked another 4 label. The aim is to find whether the content indicates self-promotion/advertisement, an attack/denouncement of opposition, or both. Here, promotion is expressed in the form of any positive sentiment, promise, or accolade expressed by the party for itself. Slogans and hashtags are often employed in such promotions. For example, SP energetically used the hashtag such as #बाइसमेंबाइसिकल (Cycle in 2022) to refer to its election symbol of a cycle as an indicator of SP turning the wheels and coming to power in 2022. On similar lines, denouncement refers to any form of negative sentiment expressed by another party. In some cases, the parties tactfully promote themselves while demoting others. Going back to the tweet in which the AAP leader talks about the kind of leaders the BJP/Congress pay homage to, he continues to state that his party will not focus on party leaders, old or new, and only pay homage to freedom fighters, therefore, qualifying as both. Hardly will there be any neutral and attacking sample, or vice versa? We observe a similar trend in Table 1 where samples marked as neutral are more likely to be either promotional or informational in nature. Whereas content marked as explicit is highly likely to be demoting/denouncing in nature. Thus, while coarse-grain annotation captures the overall sentiment, fine-grain

Table 1: Mapping annotations from phase one (L1) to phase two (L2) with the number of samples from L1 being grouped into L2. For L1, 'N, I, E' stand for the neutral, implicit, and explicit attack classes, respectively. Meanwhile, for L2, 'P, D, B, N' stands for promotion, demotion, both, and neutral.

L1/L2	P	D	B	N
N	415	13	18	249
I	15	219	92	3
E	6	517	172	1

annotation captures the intent akin to aspect-based sentiment analysis. In our case, the aspects are a pair of political parties being referred to (including self), and the sentiment is whether they are addressing the said party in a positive or negative manner. Similar to aspect-based sentiment analysis, one sentence can have multi-noun-adjective pairs, so in our case, a tweet can be both promotional and denouncing simultaneously.

Note on data curation bias. While it is impossible to fully combat biases in data curation in offense-related tasks, we took adequate measures at each pipeline step. These combined measures help annotators perform as contextually aware annotations as possible.

1. To provide a variety of voices, we looked at 111 politicians from 10 major political groups. While the content was geo-tagged to focus on Indian regions, we did not filter based on language.
2. During the annotation process, the annotators also looked at the metadata of who the speaker was and when the comment was made.
3. During three months of the elections, the annotators kept abreast with the latest election news. They employed multiple news sources for referencing while annotating samples that referred to a particular event.

Modeling. Given the limited resources for manual annotation, it is desirable to have some form of computational assistance for large-scale labeling of political attacks. We base our experiments on an n-gram model and a large language model (LLM) to achieve this. As expected, the LLM-based model (27) outperforms the indic n-gram model (28, 29), and the former is employed to annotate attack labels at scale. Appendix A.2 describes both the experimental setup and results.

Data Availability. Following Twitter's data-sharing policy, the tweet-ids, corresponding attack labels, and source code of all the analyses and modeling are available at `https://github.com/LCS2-IIITD/india_pol_attack`.

7 Statements and Declarations

The authors declare no competing interest. The authors are not affiliated with any political group. The work is not funded by any political group. S.M contributed to data collection, annotation, and modeling. S.M and T.C contributed to the analysis and paper writing.

Acknowledgements. The authors would like to thank Drishya and Priyanshi for contributing to the initial data curation and annotation. Sarah is supported by the Prime Minister Doctoral Fellowship in partnership with Wipro AI and SERB India.

References

[1] Zhuravskaya E, Petrova M, Enikolopov R. Political Effects of the Internet and Social Media. Annual Review of Economics. 2020;12(1):415-38.

[2] Jakesch M, Garimella K, Eckles D, Naaman M. Trend Alert: A Cross-Platform Organization Manipulated Twitter Trends in the Indian General Election. Proc ACM Hum-Comput Interact. 2021;5.

[3] Turbic NGB, Galesic M. Group threat, political extremity, and collective dynamics in online discussions. Scientific Reports. 2023 Feb;13(1). Available from: https://doi.org/10.1038/s41598-023-28569-1.

[4] Waller I, Anderson A. Quantifying social organization and political polarization in online platforms. Nature. 2021 Dec;600(7888):264-8. Available from: https://doi.org/10.1038/s41586-021-04167-x.

[5] Solovev K, Pröllochs N. Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity. In: Proceedings of the ACM Web Conference 2022. WWW '22. New York, NY, USA: Association for Computing Machinery; 2022. p. 3656–3661. Available from: https://doi.org/10.1145/3485447.3512261.

[6] Chakraborty T, Masud S. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. SIGWEB Newsl. 2022;2022(Winter):3:1-3:9. Available from: https://doi.org/10.1145/3522598.3522601.

[7] Kalmoe NP. Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence. Political Communication. 2014;31:545-63.

[8] Zahrah F, Nurse JRC, Goldsmith M. A Comparison of Online Hate on Reddit and 4chan: A Case Study of the 2020 US Election. In: ACM SIGAPP; 2022. p. 1797–1800.

[9] Kumar K. Changing Political Landscape in State Assembly Elections of Uttar Pradesh. Journal of Positive School Psychology. 2022:7752-62.

[10] Borah A, Singh SR. Investigating political polarization in India through the lens of Twitter. Social Network Analysis and Mining. 2022 Jul;12(1). Available from: https://doi.org/10.1007/s13278-022-00939-z.

[11] Jafri FA, Siddiqui MA, Thapa S, Rauniyar K, Naseem U, Razzak I. Uncovering the Dynamics of Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines. In: Workshop Proceedings of the 17th ICWSM; 2023. .

[12] Schmidt A, Wiegand M. A Survey on Hate Speech Detection using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media; 2017. p. 1-10.

[13] Awal MR, Cao R, Lee RK, Mitrovic S. AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. In: Karlapalem K, Cheng H, Ramakrishnan N, Agrawal RK, Reddy PK, Srivastava J, et al., editors. Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part I. vol. 12712 of Lecture Notes in Computer Science. Springer; 2021. p. 701-13. Available from: https://doi.org/10.1007/978-3-030-75762-5_55.

[14] Cao R, Lee RK, Hoang T. DeepHate: Hate Speech Detection via Multi-Faceted Text Representations. CoRR. 2021;abs/2103.11799. Available from: https://arxiv.org/abs/2103.11799.

[15] Masud S, Bedi M, Khan MA, Akhtar MS, Chakraborty T. Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '22. New York, NY, USA: Association for Computing Machinery; 2022. p. 3524–3534. Available from: https://doi.org/10.1145/3534678.3539161.

[16] Harriss J. Hindu Nationalism in Action: The Bharatiya Janata Party and Indian Politics. South Asia: Journal of South Asian Studies. 2015;38(4):712-8. Available from: https://doi.org/10.1080/00856401.2015.1089826.

[17] Siddarth D, Shankar R, Pal J. 'We do politics so we can change politics': communication strategies and practices in the Aam Aadmi Party's institutionalization process. Information, Communication & Society. 2021;24(10):1361-81. Available from: https://doi.org/10.1080/1369118X.2020.1856910.

[18] Singh AK, Jain J, Kameswari L, Kumaraguru P, Pal J. Note: Urbanization and Literacy as Factors in Politicians' Social Media Use in a Largely Rural State: Evidence from Uttar Pradesh, India. In: ACM COMPASS; 2022. p. 564–568.

[19] Piazza JA. Politician hate speech and domestic terrorism. International Interactions. 2020;46(3):431-53.

[20] Rheault L, Rayment E, Musulan A. Politicians in the line of fire: Incivility and the treatment of women on social media. Research & Politics. 2019;6(1):2053168018816228.

[21] Chan TKH, Cheung CMK, Lee ZWY. Cyberbullying on social networking sites: A literature review and future research directions. Information & Management. 2021;58(2):103411.

[22] Waseem Z, Davidson T, Warmsley D, Weber I. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In: Proceedings of the First Workshop on Abusive Language Online; 2017. p. 78-84.

[23] Petkevic V, Nai A. Political Attacks in 280 Characters or Less: A New Tool for the Automated Classification of Campaign Negativity on Social Media. American Politics Research. 2022;50(3):279-302.

[24] Nellis G. Election cycles and global religious intolerance. Proceedings of the National Academy of Sciences. 2023;120(1):e2213198120. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.2213198120.

[25] García Benítez-D'Ávila H. Populism and polarization in the digital arena: categorizing and measuring political attacks on Twitter; 2022. .

[26] Masud S, Dutta S, Makkar S, Jain C, Goyal V, Das A, et al. Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter. In: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021. IEEE; 2021. p. 504-15. Available from: https://doi.org/10.1109/ICDE51399.2021.00050.

[27] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL; 2019. p. 4171-86.

[28] Davidson T, Warmsley D, Macy M, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. In: ICWSM; 2017. p. 512-5.

[29] Kakwani D, Kunchukuttan A, Golla S, N C G, Bhattacharyya A, Khapra MM, et al. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 4948-61. Available from: https://aclanthology.org/2020.findings-emnlp.445.

A Appendix

In the supplementary, we provide details of dataset acquisition, model specifications, and additional details.

A.1 Dataset Acquisition

Within the abundance of social media platforms, we focus this study on Twitter, given (a) its rising popularity in Indian politics (11) and (b) its ease of collecting data for academic research. We configure the Twitter APIs to filter tweets based on geolocation to capture the Indian context. The curated dataset consists of $46k$ tweets amassed biweekly from January 1 – May 31, 2022, based on 111 political leaders and associates representing 18 political entities. We map different politicians to their parent party based on a politician's profile information/self-declaration, Wikipedia page, and news items (if any). These entities include political parties, syndicates, and a set of independent candidates. We term a political group as a *political party* when it officially contests in regional or national elections. On the other hand, we reserve the term *political syndicate* for a socio-religious organization that does not contest elections yet can be ideologically associated with various political parties.

Bhagwa Kranti Sena (BKS) is captured in our dataset as a syndicate. Meanwhile, the other 16 political parties included in our datasets are – Aam Aadmi Party (AAP), All India Majlis-E-Ittehadul Muslimeen (AIMIM), Apna Dal (AD), Azad Samaj Party (ASP), Bhartiya Janta party (BJP), Bahujan Samaj Party (BSP), Goa Forward Party (GFP), Indian National Congress (INC), Jammu and Kashmir Peoples Democratic Party (PDP), Lok Janshakti Party (LJP), Rashtriya Janta Dal (RJD), Rashtriya Lok Dal (RLD), Samajwadi Party (SP), Sanyukt Samaj Morcha Party (SSM), Shiromani Akali Dal (SAD) and Shiv Sena (SHS). Note that for our current analyses of political attacks, we do not differentiate between a syndicate and a party and treat all 17 as political groups capable of initiating attacks. While independent candidates are also political entities, we do not cover them in this discussion. We additionally look into the official Twitter handles of six political parties – BJP, INC, AAP, SP, BSP, and AIMIM. Overall, $46k$ posts are composed of $14,450$ non-duplicate tweets from official party handle and $32,017$ posts from politicians. To capture the diversity in interactions garnered by these parties, we look into party-wise user and tweet contributions along with average likes/retweets per tweet. Table 2 provides an overview of the same. Consequently, we notice Hindi, English, and Punjabi as the most frequent languages in our dataset. It can be understood that three out of five states where elections are held are in north India, where Hindi is the de-facto mode of communication.

A.2 Large-scale Modeling of Political Attacks

We begin with binary classification settings where the two labels are attacking and neutral. This setup combines the explicit and implicit labels into class 0, and

Table 2: The volume of posts in our dataset from the 17 political groups (Parties/Syndicates). We represent the information as the number of tweets (T), the number of unique users (U), the ratio of tweets to users (T/U), total retweets (R), average retweets per tweet (R/T), total likes (L), and, average likes per tweets (L/T). The abbreviations for the political party are – Aam Aadmi Party (AAP), All India Majlis-E-Ittehadul Muslimeen (AIMIM), Apna Dal (AD), Azad Samaj Party (ASP), Bhartiya Janta Party (BJP), Bahujan Samaj Party (BSP), Bhagwa Kranti Sena (BKS), Goa Forward Party (GFP), Indian National Congress (INC), Jammu and Kashmir Peoples Democratic Party (PDP), Lok Janshakti Party (LJP), Rashtriya Janta Dal (RJD), Rashtriya Lok Dal (RLD), Samajwadi Party (SP), Sanyukt Samaj Morcha Party (SSM), Shiromani Akali Dal (SAD) and Shiv Sena (SHS).

Name	**# T**	**# U**	**T/U**	**# R**	**R/T**	**# L**	**L/T**
AAP	3537	14	252	1447949	409	6894606	1949
AIMIM	2179	10	217	650658	298	2984679	1369
AD	214	1	214	24525	114	148145	692
ASP	140	1	140	232700	1662	969690	6926
BJP	26879	37	726	14296955	531	72566346	2699
BSP	132	3	44	197609	1497	933682	7073
BKS	720	1	720	446423	620	2426117	3369
GFP	125	1	125	2269	18	7562	60
INC	12357	33	374	5617578	454	21960713	1777
PDP	59	1	59	6819	115	39289	665
LJP	90	1	90	484	5	3411	37
RJD	86	2	43	73987	860	517034	6012
RLD	164	1	164	18320	111	120412	734
SP	6480	10	648	3551946	548	18232421	2813
SSM	12	1	12	215	17	1064	88
SAD	514	3	171	33244	64	91111	177
SHS	124	2	62	92942	749	834949	6733

neutral acts as class 1. We then extend our models for the 3-way classification with class labels 0, 1, and 2, respectively. Note that the binary classification is only performed for sanity testing as it reduces the skewness in class labels and allows us to understand if the systems are learning any pattern at all. Only the labels from the 3-way classification are employed in the labeling attack at scale. We base our experiments using two classes of models:

- **N-gram based system**: We adopt and extend the Davidson (28) model of 3-way hate speech detection. The original work builds a model to detect hate, offense, and labels in English Tweets surrounding the USA. However, given Indic languages' cultural and semantic diversity, we add language-specific features like Indic stopwords and Indic tokenization models (29).

We combine the preprocessed n-gram features and tweet-specific metadata as the final input to a logistic regression classifier. The tf-idf vectorizer and the pos-tagger vectorizer were initialized for a max feature size of 150 and min/max$_{\mathrm{df}}$ of 5/0.75. The rest of the hyperparameters of the vectorizers and logistic regression were default as present in sklearn[1].

- **LLM-based system**: Large-scale language models (LLMs) trained on texts from across the web have proven effective in various downstream tasks. One variant of LLM is mBERT (27), which is trained in 104 languages (including Indic ones). Given the small size of our dataset, we freeze the initial layers of m-BERT and then fine-tune a classification head on top of the base mBERT. For fine-tuning mBERT for maximum 120 with early stopping a batch size of 32 epochs, we employ an AdamW optimizer with a learning rate=$3e-5$, and the rest of the hyperparameters, as default in Pytorch[2].

We perform 3-fold cross-validation for training and evaluation with class weighting. To evaluate the imbalance in group truth labels, we report the macro F1-score, Precision, and Recall along with accuracy. As expected, the LLM model significantly outperforms the n-gram model. The accuracy and macro-F1 of the n-gram system stagger at 80% and 80%, respectively, for the 2-way. Meanwhile, it drops significantly during 3-way classification to 65% accuracy and 60% macro-F1. Its performance on implicit class is abysmal at 44% macro-F1. On the other hand, the LLM system achieves accuracy and macro-F1 of 98% and 98% in 2-way classification, which drops to 93% (accuracy) and 91% (macro-F1) in $3-$ way classification. Here even the implicit class enjoys a macro-F1 of 0.84%. Further, under 2-way classification, the n-gram and m-BERT models report a ROC-AUC score of 0.79 and 0.97, respectively. Meanwhile, under 3-way classification, the n-gram and m-BERT models report a ROC-AUC score of 0.71 (0.71) and 0.95 (0.95), respectively, for the OVO (OVR) multi-class setting. We observe that compared to the n-gram model, LLM based system's precision and recall values are more or less balanced (i.e., one was not much higher or lower than the other), corroborating the robustness of the system.

[1] `https://scikit-learn.org/stable/`

[2] `https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html`