International Research Journal of Management Science & Technology

ISSN 2250 – 1959(0nline)
2348 – 9367 (Print)

An Internationally Indexed Peer Reviewed & Refereed Journal

www.IRJMST.com
www.isarasolutions.com

Published by iSaRa Solutions

"I'm Just Saying": Justifying Celebrity Trolling through Moral Disengagement, Self-Reported Motivations, and Platform Affordances

Praveen[1],
Research Scholar, Department of Mass Communication, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India.

Prof. Umesh Arya[2],
Professor, Department of Mass Communication, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India.

Abstract

This study examines the self-reported motivations, moral disengagement, and platform affordances underlying the trolling of high-visibility celebrity accounts on Indian social media. The research explores, how the features of each social media platform, psychological motives, and normative justifications influence trolling behavior by employing a cross-platform survey that includes self-identified trollers from Facebook, Instagram, and Twitter/X. The affordances of the platforms are found to shape the expression of trolling, and narratives of patriotic pride, moral policing and perceived "celebritiness" of the target are common ways to rationalize trolling. Not surprisingly, given theoretical expectations, moral disengagement has a weak association with retaliatory trolling, and can be viewed as a post hoc rationalization rather than a direct behavioral motivator. The study also reveals a gap between respondents' beliefs and their reported behaviour – known as the attitude–behaviour gap. The results contribute to the moral disengagement and platform affordance theories and have implications for platform governance, online safety, and future research of online celebrity abuse..

Keywords: online trolling, moral disengagement, platform affordances, celebrity harassment, social media, India, gendered online abuse, survey research

Introduction

Online trolling has become a significant and ubiquitous form of interpersonal, political and para-social aggression on today's social networking sites (SNS). The conditions in India, where the social media user base has grown at an unprecedented rate, the personality-centric 'culture of celebrities' in the realms of cinema, cricket and politics, and the polarized political climate have led to a situation where, for a significant minority of users, trolling of popular public personalities is not only ubiquitous but also normatively acceptable. It is this population that this study is seeking to focus on: people that have identified as trolls of the most followed celebrity accounts on Facebook, Instagram, and Twitter/X in 2020, a period when people have been migrating to social media platforms, facing increased polarization, and seeing numerous online mobilization events causing outrage and collective anger (such as #JusticeForSSR, as well as organized boycott hashtags that repeatedly targeted celebrity accounts as points of coordinated hostility).

Trolling is here defined as "the intentional posting of content that is provocative, inflammatory, disruptive or deceptive within a community, with the aim of provoking an emotional response, disrupting the flow of communication, and/or claiming to be superior to the target, whether or not the troller is sincere regarding the content of that posting" (Buckels et al., 2014, p. 370). This is meant to be differentiated from cyberbullying, which tends to be dyadic, repeated and imbalanced, and from flaming, a one-off hostile exchange. As used here, "trolling" is a performative act with an audience, usually done in the obvious comment section of celebrities' accounts, where the audience is guaranteed, and thus the chances of response are high.

It is not a uniform phenomenon. It can be typologised in four overlapping ways: (1) troll type – abusive or political or sarcastic; (2) strategic approach – provocation, misinformation or ridicule; (3) affective tone – aggressive or mocking; (4) target orientation – the person of the celebrity, or the ideological position that he/she is believed to represent. They do not work in isolation, as this will be important because policy, moderation, and psychological interventions will not be the same for ideologically motivated abuse and entertainment abuse.

Several trends underline the growing relevance of this inquiry. India's social media user base ran into several hundred million active accounts over the study period, with celebrity accounts among the most-followed globally, concentrating a large pool of potential trollers around a small number of accounts (DataReportal, 2021). The 2020 ban on TikTok accelerated adoption of Instagram Reels and Stories, shifting the platform's idiom of hostility toward image-based mockery (The Indian Express, 2020). Coordinated hashtag activity around the 2019 general election and subsequent controversies showed that celebrity-directed trolling is often organised and network-mediated rather than purely spontaneous (Gupta et al., 2020; Pal et al., 2019). Finally, documented asymmetries in abuse directed at female versus male public figures have drawn sustained attention from rights monitors, raising the empirical question of whether this reflects individual-level gender differences in trolling propensity or structural targeting logics independent of who is doing the trolling (Amnesty International, 2020; IT for Change, 2022).

Here are three platforms each representing different interaction architectures. Facebook is a place for lengthy, antagonistic, continued dialogue; Twitter/X is a place for brief, high-visibility, often threatening dialogue; Instagram, visual and youth-favored, is a place for trolling in captions and meme appropriation. When the population of interest is known, but the technologies vary, a cross-platform design provides analytic leverage to isolate individual disposition from the technology that influences its expression.

Given this, the current study examines 1,106 self-identified frequent trollers of the 10 most-followed celebrity Facebook, Instagram, and Twitter/X accounts for each service (as of 2020), using both demographic profiling and Likert-scale and categorical analyses of trolling behaviours, trolling norms, and community reactions. The rest of this paper examines the related literature, outlines theory, research questions, hypotheses and purposes, explains sampling process and methodology, reports findings and discusses results, and concludes with terminologies and limitations which constrained their interpretations.

Literature Review

Research on online trolling can be divided into three strands: individual-differences psychology (who trolls and why), sociotechnical affordance theory (how platform design contributes to hostility), and political-communication research (trolling as co-ordinated mobilisation). All three of them are answered by the dataset on which this study is based.

Individual Differences and Moral Disengagement

Research on online trolling can be divided into three strands: individual-differences psychology (who trolls and why), sociotechnical affordance theory (how platform design contributes to hostility), and political-communication research (trolling as co-ordinated mobilisation). All three of them are answered by the dataset on which this study is based.

Platform Affordances and Comparative Behaviour

boyd's (2010) concept of networked publics holds that a platform's persistence, replicability, scalability, and searchability condition the form antagonism takes without necessarily changing its underlying prevalence. Indian studies document that Facebook's threaded architecture sustains community-level polarisation (Bhushan, 2015), that Twitter's brevity and visibility favour real-time political mobilisation (Pal et al., 2019; Sharma & Sivakumar, 2023), and that Instagram's visual, youth-skewed character has repositioned meme-based ridicule as its dominant idiom post-2020 (The Indian Express, 2020). This literature relies overwhelmingly on trace analysis rather than trollers' own self-reported cognitions. The SIDE model (Postmes, Spears, & Lea, 1998) bridges the affordance and disinhibition accounts by proposing that anonymity amplifies conformity to in-group norms rather than producing generic disinhibition, consistent with the coordinated hostility documented on Indian Twitter/X (Chaudhuri, 2021); Cheng, Danescu-Niculescu-Mizil, and Leskovec (2017) similarly found trolling probability rises sharply after exposure to a hostile thread, regardless of baseline disposition. Pew data on platform-specific user composition (Anderson & Jiang, 2018; Wojcik & Hughes, 2019) further substantiate the demographic skew this study's own sample reproduces.

Coordinated and Politically Motivated Trolling in India

A third strand documents the organised character of celebrity- and politician-directed trolling in India, including coordinated hashtag activity during the 2019 election (Gupta et al., 2020) and gendered, sexualised abuse disproportionately targeting women in public life (Amnesty International, 2020; IT for Change, 2022). This raises an unresolved tension motivating one of this study's central objectives: if gendered targeting reflects coordinated, norm-enforced campaigns rather than individual male hostility, self-report data from trollers themselves should show minimal gender differences even as trace data shows large targeting asymmetries, a pattern consistent with international findings that gendered harassment functions as a networked, norm-reinforced practice (Citron, 2014; Lewis, Rowe, & Wiper, 2017; Vitak, Chadha, Steiner, & Ashktorab, 2017). Udupa (2019) situates this within an Indian "gaali" (abuse) register that normalises ideologically framed hostility as legitimate speech, while related work on digital Hindu-nationalist publics (Udupa, 2018) documents the ideological scaffolding respondents draw on to rationalise trolling as civic

participation. Because this study's targets are celebrities, the parasocial and anti-fan literatures are also relevant: Giles (2002) frames parasocial attachment as curdling into hostility upon perceived violation, and Click and Scott (2018) document organised anti-fandom mobilising ridicule in terms analogous to the deservingness justifications measured here.

Identified Gaps

Three gaps justify this study's objectives. Methodologically, little Indian research has administered a harmonised survey simultaneously across three structurally distinct platforms to the same population, precluding rigorous cross-platform statistical comparison. Empirically, normative justification and psychological motive have been treated as narrative themes rather than statistically testable constructs linked to concrete behavioural indicators. Theoretically, the gender-asymmetry literature documents targeting outcomes but rarely tests, with self-report data from trollers themselves, whether individual-level psychological profiles differ enough by gender to account for those outcomes. The framework, objectives, and hypotheses below are designed to close these gaps.

Theoretical Framework, Objectives, Research Questions, Hypotheses, and Significance

Theoretical Framework

This study draws on three complementary theoretical models. Bandura's (1999) theory of moral disengagement anchors the normative-justification and guiltlessness measures, moral justification, euphemistic language, diffusion of responsibility. boyd's (2010) affordance theory of networked publics explains cross-platform differences as products of persistence, scalability, and searchability rather than differing user psychology. Together these generate a testable proposition: trolling propensity is a stable disposition, expressed via online disinhibition and rationalised via moral disengagement, whose surface form, but not underlying cause, is shaped by platform affordance. As Section 6 (Results and Discussion) shows, the absence of association between the disengagement composite and retaliation is unexpected, and suggests disengagement functions here as a post hoc justificatory narrative rather than a proximal behavioural driver.

Objectives

Broad Objective: To empirically examine, across Facebook, Instagram, and Twitter/X, the behavioural characteristics, normative justifications, and psychological motives of Indian users who report frequently trolling high-visibility celebrity accounts, and to assess how these self-reports converge with, or diverge from, platform, gender, and age differences documented in the literature.

- Objective 1: Characterise the behavioural profile of celebrity-directed trolling (tone, strategy, disruptive tactics) and test whether it differs significantly across platforms.
- Objective 2: Examine the normative and ideological frames, nationalism, moral policing, perceived deservingness, through which respondents justify trolling.
- Objective 3: Investigate the psychological motives underlying trolling, entertainment, retaliation, moral disengagement, and test whether these are patterned by gender and age.
- Objective 4: Assess perceptions of community response to trolling and the consistency between stated beliefs about engagement and respondents' own retaliatory conduct.

Research Questions

- RQ1a/b: How does self-reported trolling tone (mocking language, threats, off-topic disruption) differ across platforms, and which differences are statistically significant?
- RQ2a/b: What proportion of respondents endorse nationalist or moral-policing justification, does this vary by platform or age, and does it correlate with concrete behavioural indicators?
- RQ3a-c: Is moral disengagement associated with retaliation; do trolling indicators differ by gender or age?
- RQ4a/b: What community responses do respondents perceive as most common, and is there a gap between believing engagement worsens a situation and personally retaliating?

Hypotheses

The hypotheses below are stated with the empirical outcome observed, consistent with reporting disconfirming results plainly rather than smoothing them into the expected narrative. Full statistics appear in the Results and Discussion section and its tables.

- H1: Trolling tone differs significantly by platform (Instagram highest mocking, Twitter/X highest threats). Outcome: supported (mocking, $F(2,1103)=3.58$, $p=.028$; threats, $F(2,1103)=8.61$, $p<.001$).
- H2: Post-conflict energization is uniformly moderate-to-high across platforms. Outcome: partially supported — means cluster narrowly (3.43–3.54) yet the platform effect remains significant ($F(2,1103)=4.80$, $p=.008$).
- H3: Patriotic justification is endorsed by a substantial minority and more prevalent among younger users. Outcome: partially supported — 42.8% overall endorsement with a descriptive age decline that did not reach significance ($\chi^2(6)=9.50$, $p=.147$).
- H4: Moral disengagement positively predicts retaliation. Outcome: not supported — negligible, non-significant correlation ($r=-.014$, $p=.633$).
- H5: Gender does not significantly differentiate trolling propensity. Outcome: supported — all ten indicators non-significant ($p>.13$; $|d|<0.12$).
- H6: An attitude–behaviour gap exists between believing engagement worsens a situation and personally retaliating. Outcome: supported — 51.7% agreement versus 41.8% self-reported retaliation, with no significant platform difference ($\chi^2(2)=1.93$, $p=.380$).

Significance of the Study

This study contributes to empirical knowledge in three ways. First, it offers one of few statistically comparable, three-platform Indian survey datasets, enabling ANOVA and chi-square testing unavailable to single-platform content analysis. Second, by testing, and failing to confirm, the assumed link between moral disengagement and retaliation, it complicates a causal chain the trolling literature often takes for granted. Third, by showing negligible individual-level gender differences in trolling propensity alongside large documented targeting asymmetries, it lends direct empirical support to a structural, norm-enforced account of gendered online abuse, with implications for moderation policy that would otherwise misattribute the problem to individual disposition.

Sampling and Methodology

This study adopts a quantitative, cross-sectional survey design suited to statistically comparing self-reported trolling behaviour across three platforms and demographic strata within a bounded reference

period. A structured, self-administered questionnaire permits standardised measurement of otherwise unobservable states, justification, guilt, motive, that trace-based content analysis cannot directly capture.

Universe of the Study

The universe comprises all adult Indian social media users active on Facebook, Instagram, or Twitter/X during 2020 who had posted trolling comments on the accounts of the ten most-followed public figures on each platform (identified via triangulated SocialBlade, Sacnilk, and Emplifi follower-ranking data; Emplifi, 2020; SocialBlade, 2021). Because India's combined active user base across these platforms ran into several hundred million accounts (DataReportal, 2021), and no platform discloses who has specifically trolled a given account, this universe is unenumerable in the conventional census sense and is defined intensionally by behavioural and platform criteria rather than as a finite, listable population.

Table 1. Selected High-Visibility Celebrity Accounts (2020)

Platform	Top 10 Verified Accounts (2020)
Instagram	Virat Kohli (@virat.kohli), Priyanka Chopra (@priyankachopra), Deepika Padukone (@deepikapadukone), Alia Bhatt (@aliaabhatt), Akshay Kumar (@akshaykumar), Neha Kakkar (@nehakakkar), Shahid Kapoor (@shahidkapoor), Shraddha Kapoor (@shraddhakapoor), Jacqueline Fernandez (@jacquelinef143), Salman Khan (@beingsalmankhan)
Facebook	Narendra Modi, Virat Kohli, Deepika Padukone, Priyanka Chopra, Salman Khan, Shah Rukh Khan, Aamir Khan, Amitabh Bachchan, Hrithik Roshan, Sachin Tendulkar
Twitter	Narendra Modi (@narendramodi), Amitabh Bachchan (@SrBachchan), Shah Rukh Khan (@iamsrk), Akshay Kumar (@akshaykumar), Sachin Tendulkar (@sachinrt), Salman Khan (@BeingSalmanKhan), Virat Kohli (@imVkohli), Deepika Padukone (@deepikapadukone), Hrithik Roshan (@iHrithik), Priyanka Chopra (@priyankachopra)

Sample and Sampling Technique

A non-probability purposive-cum-snowball strategy was used, since frequent trollers of specific accounts are not a listed or randomly sampleable population. Initial respondents were identified through platform-based outreach targeted at visibly hostile commenters, then asked to refer similarly engaged users, following standard online snowball recruitment for hidden populations (Baltar & Brunet, 2012). Recruitment continued until quota targets approximating each platform's share of identified troll-engagement volume were reached. This approach is standard for populations with no sampling frame (Etikan, Musa, & Alkassim, 2016), but limits generalisability beyond the sample, a constraint reflected in the total survey error framework's caution that non-random recruitment trades representativeness for access (Groves et al., 2009).

Sample Size

The realised sample comprised 1,106 respondents: 346 from Twitter/X (31.3%), 385 from Facebook (34.8%), and 375 from Instagram (33.9%). Each sub-sample exceeds $n = 385$, the minimum required

by Cochran's formula for stable proportion estimation at 95% confidence and a 5% margin of error, and provides adequate power (> .80 at $\alpha = .05$) to detect small-to-medium effects in the tests reported in the Results and Discussion section, following Cohen's (1988) conventions.

Time Frame of Data Collection

Two time frames are relevant and should not be conflated. The behavioural reference period, the period respondents were asked to report on, is 2020, when the sampled accounts' follower rankings were established and pandemic-driven mobilisation intensified celebrity-directed trolling. The data-collection window ran from January to December 2022 for Facebook and Instagram, with Twitter/X responses extending in a few cases into early January 2023. Respondents were thus surveyed retrospectively, one-and-a-half to three years after the reference period, a lag whose implications for recall accuracy are addressed under Limitations. The reference period was uneven because of the nature of push of questionnaire on different platforms. Separate URL of Google Forms questionnaire were pushed to messaging inbox of profile whom published a troll comment on chosen profiles.

Tools and Unit of Analysis

Data were collected via a Google Forms questionnaire sent to individual person who engaged in trolling on profiles selected, and then exported to CSV and analysed in Python (pandas; SciPy for one-way ANOVA, independent-samples t-tests, chi-square tests, and point-biserial correlation) alongside Excel for cross-tabulation checks. The survey used closed Likert and categorical items only, so the unit of analysis is the individual respondent ($n = 1{,}106$) rather than an individual comment or post.

Results and Discussion

The results below are organised by the four objectives outlined earlier, with each table followed by a discussion of its bearing on the hypotheses and the theoretical framework guiding this study.

Sample Profile

Table 2. Demographic Composition of the Survey Sample by Platform (N = 1,106)

Demographic	Twitter (%)	Facebook (%)	Instagram (%)	Overall (%)
Age 18–25	21.1	31.4	51.5	35.0
Age 26–35	50.6	40.8	32.5	41.0
Age 36–45	19.4	19.2	11.7	16.7
Age 45+	9.0	8.6	4.3	7.2
Male	73.4	67.8	65.1	68.6
Female	24.6	30.1	33.1	29.4
Urban location	53.2	53.0	50.1	52.1
Rural location	32.9	35.6	35.7	34.8
Graduate education	53.8	47.5	49.6	50.2
Social media 3–5 hrs/day	37.9	48.3	41.9	42.9

Table 2 shows a male-skewed sample (68.6%) concentrated in the 18–35 age range (76.0% combined), with a near-even urban/rural split and roughly half holding a graduate qualification. Twitter/X respondents were comparatively older and more male-skewed; Instagram respondents were markedly younger (51.5% aged 18–25) with the highest share of female respondents (33.1%), corroborating Pew findings on platform-specific age and gender composition (Anderson & Jiang, 2018; Wojcik & Hughes, 2019). This profile serves as contextual background for the behavioural results that follow.

Platform Differences in Trolling Behaviour (Objective 1)

Table 3. Mean Self-Reported Trolling Behaviour Indicators by Platform (1–5 Likert Scale) and One-Way ANOVA

Indicator	Twitter (M)	Facebook (M)	Instagram (M)	F(2,1103)	p
Mocking language frequency	3.05	3.04	3.22	3.58	.028*
Enjoy upsetting others	2.92	2.99	2.94	0.66	.518
Use threats	2.09	2.04	1.78	8.61	<.001***
Energized after conflicts	3.43	3.52	3.54	4.80	.008**
Moral policing necessary	3.00	2.98	3.01	0.09	.916
Manipulate others for fun	3.51	3.49	3.50	0.12	.887
No guilt for harsh comments	3.50	3.53	3.54	0.47	.624

Table 3 depicts significant platform differences in mocking language, threat use, and post-conflict energisation, but not in enjoyment of upsetting others, moral policing, manipulation, or guiltlessness. Instagram recorded the highest mocking mean and lowest threat mean; Twitter/X recorded the highest threat mean, directly answering RQ1a/b and supporting H1: Instagram's visual, meme-driven register plausibly channels hostility toward ridicule rather than direct threat (The Indian Express, 2020), while Twitter/X's brevity and political amplification favour blunter exchange (Pal et al., 2019; Sharma & Sivakumar, 2023). The energisation result partially supports H2: means cluster narrowly across platforms, yet the effect remains significant, indicating a broadly shared but not fully platform-invariant entertainment motive.

Normative Justification and Behavioural Indicators (Objective 2)

Table 4. Categorical Trolling Behaviours and Attitudes, Overall Endorsement and Platform Chi-Square

Behaviour / Attitude	Overall (%)	χ^2 (df)	p
Post off-topic content to derail	38.6	7.96 (2)	.019*
Troll for national pride/patriotism (Yes)	42.8	3.37 (4)	.498
Use hashtags specifically for attacks	27.3	3.01 (2)	.222
Believe celebrities deserve criticism	62.5	0.70 (2)	.705
Retaliate when trolled	41.8	1.93 (2)	.380
Fans defend celebrities against trolls	67.2	7.53 (2)	.023*
Joined mass-report/moderation group	37.8	0.57 (2)	.751
Believe engagement worsens situation: Disagree / Neutral / Agree	33.0 / 33.1 / 33.9	1.69 (4)	.793

The upper rows of Table 4 address Objective 2: patriotic justification was endorsed by a substantial minority (42.8%), partially supporting H3, though its descriptive decline with age was not significant ($\chi^2(6) = 9.50$, $p = .147$), so this should be read as suggestive only (RQ2a). Belief that celebrities generically "deserve" criticism was both highly endorsed (62.5%) and statistically indistinguishable across platforms, while off-topic derailment differed significantly by platform ($\chi^2 = 7.96$, $p = .019$) but coordinated hashtag attacks did not (RQ2b). Views on whether engaging with trolls worsens a situation were close to evenly split three ways — roughly a third disagree, a third are neutral, and a third agree — with no significant platform difference ($\chi^2(4) = 1.69$, $p = .793$). Read together, moral-policing and deservingness justification appear to form a broadly shared, platform-independent belief system, consistent with Udupa's (2019) account of a normalised abuse register, whereas organised behaviours such as off-topic derailment retain a platform-specific signature likely tied to Facebook's thread-hijacking affordance.

Psychological Motives and Demographic Patterning (Objective 3)

Table 5. Gender and Age Comparisons on Key Self-Report Indicators

Test	Statistic	p	Effect size
Gender: Use threats (M vs. F)	t = 1.08	.279	d = 0.070
Gender: No guilt (M vs. F)	t = -0.85	.395	d = -0.056
Gender: Manipulate for fun (M vs. F)	t = 1.51	.131	d = 0.100
Age: Mocking language (ANOVA)	F(3,1102) = 1.20	.308	—
Age: Retaliation (χ^2, df = 3)	0.46	.928	—
Disengagement composite × Retaliation (r)	r = -0.014	.633	negligible

Table 5 shows a negligible, non-significant correlation between the moral-disengagement composite and retaliation ($r = -.014$, $p = .633$), disconfirming H4 (RQ3a) and suggesting disengagement functions here as retrospective self-justification rather than a proximal trigger for retaliatory acts. No

gender differences reach significance across any indicator (all $p > .13$; $|d| < 0.12$), supporting H5 (RQ3b): because male and female trollers report statistically indistinguishable hostility, the well-documented gender asymmetry in who is targeted (Amnesty International, 2020; IT for Change, 2022) is more plausibly explained by collectively enforced targeting norms (Citron, 2014; Lewis, Rowe, & Wiper, 2017; Vitak, Chadha, Steiner, & Ashktorab, 2017) than by differing individual disposition. Age likewise fails to structure either mocking language or retaliation significantly (RQ3c), indicating that, unlike platform, age does not meaningfully pattern trolling propensity once patriotic justification is set aside.

Community Response and the Attitude–Behaviour Gap (Objective 4)

The lower rows of Table 4 address Objective 4. Fan defence of celebrities was the most commonly perceived community response (67.2%), well ahead of joining a mass-report group (37.8%) or receiving support from strangers (17.2%), suggesting informal fan mobilisation is perceived as a stronger counter-force than formal moderation (RQ4a). A clear attitude–behaviour gap emerged for RQ4b and H6: 51.7% agreed at some level that engaging with trolls worsens a situation, yet only 41.8% reported personally retaliating, a gap that did not vary significantly by platform ($\chi^2(2) = 1.93$, $p = .380$), indicating the discrepancy is a general rather than platform-specific feature of the sample.

A closer look at how respondents describe their own coping response sharpens this picture. Cross-tabulating respondents' stated reaction to being trolled (ignore, report, or engage) against their separate report of retaliation shows the same inconsistency reappearing within behavioural self-reports themselves, not only between belief and conduct: 40.9% of respondents who say they typically ignore trolls nonetheless report retaliating, compared with 37.9% of those who say they engage, and 45.1% of those who say they report the trolling. Table 6 presents this comparison.

Table 6. Reported Reaction to Trolls Cross-Tabulated with Reported Retaliation

Reported reaction to trolls	n	Retaliate: No (%)	Retaliate: Yes (%)
Ignore	599	59.1	40.9
Report	346	54.9	45.1
Engage	161	62.1	37.9

Respondents who describe themselves as typically ignoring or reporting trolls are no less likely to report retaliating than those who describe themselves as engaging — if anything, slightly more likely. This extends the attitude–behaviour gap identified above: the dissociation is not confined to the distance between an abstract belief and a single behavioural report, but also appears between two behavioural self-reports that describe the same underlying conduct from different angles, which points toward a genuine instability in how respondents categorise their own actions after the fact rather than a simple mismatch between principle and practice.

Synthesis

Overall, platform affordance exerts a meaningful influence on the register of hostility — tone, threat use, off-topic derailment — but not on the justificatory belief system or the demographic patterning of who trolls, to the extent that boyd's (2010) affordance theory coexists with, but does not supplant, a stable disposition rationalised through moral disengagement (Bandura, 1999). Where a platform's design plausibly channels expression, the three platforms differ significantly; where the outcome is closer to a belief or a disposition — moral-policing endorsement, guiltlessness, manipulation for its own sake, gender, age — they mostly do not.

The most theoretically significant result of the study is that disengagement is not meaningfully correlated with retaliatory behaviour, thereby challenging the causal chain presumed in much of the literature and suggesting that disengagement represents primarily a retrospective process of self-justification. The attitude–behaviour evidence reinforces this reading from a second angle: respondents' stated beliefs about engagement, their reported coping style, and their reported retaliation do not line up consistently with one another, which is easier to explain if disengagement operates as an account people give afterward than as a mechanism that drives the behaviour in the first place.

None of this is contradicted by the demographic profile of the sample, which remains young and male-skewed (76.0% aged 18–35) with no meaningful gender difference across any of the ten trolling indicators tested. Set against the scale of documented targeting asymmetries by gender in Indian public life (Amnesty International, 2020; IT for Change, 2022), a sample of self-identified trollers reporting no meaningful gender difference in their own hostility sits more comfortably with a norm-enforced, structural account of who gets targeted than with one resting on differing individual disposition.

Conclusion

This study investigated the intersection of platform affordances, normative justification, and psychological motive in trolling high-visibility celebrity accounts on Indian social media by conducting a harmonised survey with 1,106 self-identified frequent trollers. The architecture of the platform has a meaningful effect on the expressive register of hostility — sharper mocking on Instagram, threat-laden brevity on Twitter/X, extended antagonism and off-topic derailment on Facebook — and leaves the underlying belief system and demographic patterning of who trolls relatively unmoved. Respondents justify their behaviour mainly in terms of patriotism, moral policing, and perceived celebrity deservingness, all rooted in a broader ethical framework and consistently endorsed across platforms and age groups.

Two findings matter most for the study's contribution. First, moral disengagement shows no meaningful association with retaliation, challenging an often-assumed causal chain in this research and suggesting instead that moral disengagement functions as retrospective self-justification. Second, there is no appreciable difference in trolling propensity between male and female respondents, which, read against well-established targeting asymmetries, lends empirical weight to a structural, norm-enforced explanation of gendered online abuse over an individual-disposition explanation. A closer look at respondents' own coping reports adds a further layer to this picture: the gap between

what respondents say they believe and what they say they do extends into comparisons between different behavioural self-reports as well, suggesting a broader instability in how this population narrates its own conduct rather than a narrow gap between one belief and one behaviour.

The results indicate that trolling is unlikely to be substantially reduced by adjusting platform-specific technical affordances alone — which is not to say such interventions are unnecessary — and that coordinated targeting norms and fan or anti-fan mobilisation may matter more than targeting individual trollers. Future research should combine this self-report measure with matched behavioural trace data to determine whether the dissociation between disengagement and retaliation reflects a genuine absence of causal influence or a shared-method artefact, and should broaden the comparative design to a probability-based sample to strengthen generalisability to a wider population.

Terminology and Glossary

- Trolling: Deliberate posting of provocative, disruptive, or deceptive content intended to elicit an emotional reaction, disrupt discourse, or assert dominance over a target, regardless of the poster's sincere beliefs (Buckels et al., 2014; Hardaker, 2010).
- Moral disengagement: Cognitive mechanisms, moral justification, euphemistic labelling, diffusion of responsibility, disregard of consequences, through which individuals neutralise self-censure for harmful conduct (Bandura, 1999); operationalised here as guiltlessness, enjoyment of manipulation, and enjoyment of others' distress (Runions & Bak, 2015).
- Online disinhibition effect: Reduced social restraint online attributable to anonymity, invisibility, asynchronicity, and dissociation of online persona from offline identity (Suler, 2004).
- Platform affordance: The technical and social possibilities a platform's design makes available, persistence, replicability, scalability, searchability, shaping without wholly determining the form of behaviour expressed on it (boyd, 2010).
- Normative justification: A respondent's rationale for trolling framed as adherence to a legitimate norm, nationalism or moral policing, rather than as gratuitous harassment (Udupa, 2019; Postmes, Spears, & Lea, 1998).
- Celebrity deservingness: The belief that a public figure's visibility or conduct forfeits protection from harsh commentary, blurring the line between criticism and harassment (Click & Scott, 2018; Giles, 2002).
- Retaliatory trolling: Trolling reported as a reactive response to being trolled or criticised, distinct from unprovoked trolling (Cheng, Danescu-Niculescu-Mizil, & Leskovec, 2017).
- Attitude–behaviour gap: A measurable discrepancy between a stated belief (engagement worsens a situation) and reported conduct (continuing to retaliate), consistent with disengagement permitting belief and behaviour to remain dissociated (Bandura, 1999).
- Structural (norm-enforced) targeting: An account in which targeting asymmetries (e.g., by gender) arise from collective norms about legitimate targets rather than individual-level differences among trollers (Citron, 2014; Lewis, Rowe, & Wiper, 2017; Vitak, Chadha, Steiner, & Ashktorab, 2017).
- Frequent troller: This study's operational population — a self-identifying respondent reporting repeated trolling of top-followed celebrity accounts on the surveyed platform (Buckels et al., 2014; Cheng et al., 2017).

Limitations

- Recall and social-desirability bias: The survey was administered retrospectively (mostly 2022, with a few Twitter/X responses into early 2023) about 2020-era behaviour, exposing responses to recall decay. Self-reported threat use (M=1.78–2.09) exceeds typical trace-level prevalence, suggesting some over- or under-reporting relative to actual conduct.
- Non-probability sampling and generalisability: No enumerable sampling frame of celebrity-account trollers exists, so the purposive/snowball strategy, while appropriate for this hard-to-sample population, cannot guarantee representativeness; results describe this sample rather than a generalisable population without further replication.
- Self-report-only measurement: All behavioural indicators, including retaliation, are self-reports rather than verified trace data, so the null disengagement–retaliation association may reflect a genuine absence of relationship, an imprecise operationalisation, or shared-method artefacts; future work pairing this instrument with behavioural trace data is needed to adjudicate.

References

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). https://doi.org/10.1037/0000165-000

Amnesty International. (2020). Troll Patrol India: Exposing online abuse faced by women politicians in India. Amnesty International.

Anderson, M., & Jiang, J. (2018). Teens, social media and technology 2018. Pew Research Center.

Baltar, F., & Brunet, I. (2012). Social research 2.0: Virtual snowball sampling method using Facebook. Internet Research, 22(1), 57–74. https://doi.org/10.1108/10662241211199960

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review, 3(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Bhushan, B. (2015). Digital polarisation and the Indian public sphere: Facebook, community, and conflict. Economic and Political Weekly, 50(29), 45–52.

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), A networked self: Identity, community, and culture on social network sites (pp. 39–58). Routledge.

Bradshaw, S., & Howard, P. N. (2019). The global disinformation order: 2019 global inventory of organised social media manipulation. Oxford Internet Institute.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. Personality and Individual Differences, 67, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Chaudhuri, M. (2021). IT cells, digital armies, and the manufacture of political consensus in India. South Asian Journal of Communication, 11(1), 44–61.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing, 1217–1230. https://doi.org/10.1145/2998181.2998213

Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press.

Click, M. A., & Scott, S. (2018). The Routledge companion to media fandom. Routledge.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Craker, N., & March, E. (2016). The dark side of Facebook: The Dark Tetrad, negative social potency, and trolling behaviours. Personality and Individual Differences, 102, 79–84. https://doi.org/10.1016/j.paid.2016.06.043

Dash, S., Chakraborty, C., Giri, S. K., & Roy, S. (2022). Social media influence on Indian political discourse: A network analysis. Journal of Information Technology & Politics, 19(2), 178–195.

DataReportal. (2021). Digital 2021: India. DataReportal. https://datareportal.com/reports/digital-2021-india

Emplifi (formerly Socialbakers). (2020). India social media benchmark report 2020. Emplifi.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. American Journal of Theoretical and Applied Statistics, 5(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Field, A. (2013). Discovering statistics using IBM SPSS statistics (4th ed.). Sage.

Giles, D. C. (2002). Parasocial interaction: A review of the literature and a model for future research. Media Psychology, 4(3), 279–305. https://doi.org/10.1207/S1532785XMEP0403_04

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (2nd ed.). Wiley.

Gupta, S., Singh, N., & Kumar, R. (2020). Hashtag politics: Coordinated messaging and digital campaigning during the 2019 Indian general election. Media, Culture & Society, 42(7–8), 1310–1329.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. Journal of Politeness Research, 6(2), 215–242. https://doi.org/10.1515/jplr.2010.011

The Indian Express. (2020, July 2). TikTok ban: How Reels, Moj and other apps are filling the gap. The Indian Express.

IT for Change. (2022). Automating the mob: Gendered disinformation and trolling in South Asian digital spaces. IT for Change.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Computers in Human Behavior, 28(2), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

Lewis, R., Rowe, M., & Wiper, C. (2017). Online abuse of feminists as an emerging form of violence against women and girls. British Journal of Criminology, 57(6), 1462–1481. https://doi.org/10.1093/bjc/azw073

March, E. (2019). Psychopathy, sadism, empathy, and the motivation to cause harm: New evidence confirms malevolent nature of the Internet Troll. Personality and Individual Differences, 141, 133–137. https://doi.org/10.1016/j.paid.2019.01.001

Pal, J., Chandra, P., & Vydiswaran, V. G. V. (2019). Twitter and the Indian political landscape: Public discourse in the 2019 Lok Sabha election. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 175.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. Communication Research, 25(6), 689–715. https://doi.org/10.1177/009365098025006006

Runions, K. C., & Bak, M. (2015). Online moral disengagement, cyberbullying, and cyber-aggression. Cyberpsychology, Behavior, and Social Networking, 18(7), 400–405. https://doi.org/10.1089/cyber.2014.0670

Sest, N., & March, E. (2017). Constructing the cyber-troll: Psychopathy, sadism, and empathy. Personality and Individual Differences, 119, 69–72. https://doi.org/10.1016/j.paid.2017.06.038

Sharma, A., & Sivakumar, K. (2023). Real-time political mobilisation on Twitter/X in the Indian context: A platform affordance analysis. New Media & Society, 25(4), 812–831.

SocialBlade. (2021). India social media rankings and analytics [Data set]. SocialBlade. https://socialblade.com

Suler, J. (2004). The online disinhibition effect. Cyberpsychology & Behavior, 7(3), 321–326. https://doi.org/10.1089/1094931041291295

Udupa, S. (2018). Enterprise Hindutva and social media in urban India. Contemporary South Asia, 26(4), 453–467. https://doi.org/10.1080/09584935.2018.1544902

Udupa, S. (2019). Gaali cultures: The politics of abusive exchange on social media. New Media & Society, 21(9), 1506–1522. https://doi.org/10.1177/1461444818821376

Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017). Identifying women's experiences with and strategies for mitigating negative effects of online harassment. Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing, 1231–1245. https://doi.org/10.1145/2998181.2998337

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. Pew Research Center.

Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), Nebraska symposium on motivation (Vol. 17, pp. 237–307). University of Nebraska Press.

भारतीय भाषा, शिक्षा, साहित्य एवं शोध

ISSN 2321 – 9726

WWW.BHARTIYASHODH.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SCIENCE & TECHNOLOGY

ISSN – 2250 – 1959 (0) 2348 – 9367 (P)

WWW.IRJMST.COM

INTERNATIONAL RESEARCH JOURNAL OF COMMERCE, ARTS AND SCIENCE

ISSN 2319 – 9202

WWW.CASIRJ.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SOCIOLOGY & HUMANITIES

ISSN 2277 – 9809 (0) 2348 - 9359 (P)

WWW.IRJMSH.COM

INTERNATIONAL RESEARCH JOURNAL OF SCIENCE ENGINEERING AND TECHNOLOGY

ISSN 2454-3195 (online)

WWW.RJSET.COM

INTEGRATED RESEARCH JOURNAL OF MANAGEMENT, SCIENCE AND INNOVATION

ISSN 2582-5445

WWW.IRJMSI.COM

JOURNAL OF LEGAL STUDIES,POLITICS AND ECONOMICS RESEARCH

WWW.JLPER.COM