Uneven Toxicity Dynamics: A Multi-Dimensional Analysis of Toxicity and Engagement in Instagram News

Sehrish Qummar
Linköping University
Sweden
sehrish.qummar@liu.se

Alireza Mohammadinodooshan
Linköping University
Sweden
alireza.mohammadinodooshan@liu.se

Niklas Carlsson
Linköping University
Sweden
niklas.carlsson@liu.se

Abstract

Despite Instagram's growing role in news consumption, little is known about how engagement patterns, toxic language, and audience responses vary across news publishers. We address this gap through a comprehensive analysis of 284,843 posts from 1,026 U.S. news publishers collected over six months, characterizing outlets by *political bias* and *factual reporting quality* to examine how these attributes systematically relate to engagement patterns, toxicity usage, and audience reactions. Our results show that engagement on Instagram systematically favors biased and lower–reporting-quality outlets, which also produce substantially more toxic content. Crucially, toxicity functions as an engagement amplifier: outlets with high baseline engagement—particularly low-reporting and right-leaning publishers—capture a disproportionate share of interaction on toxic posts, whereas high-reporting and centrist outlets do not. Extending beyond posts, our analysis of over 145,257 comments shows that toxic posts elicit significantly more toxic responses, with the strongest escalation among right-leaning audiences. This effect persists across reporting levels and comment stances, suggesting that toxicity functions also as in-group reinforcement rather than solely as out-group hostility. Together, these findings provide the first large-scale evidence of how toxicity, engagement, and audience response interact within Instagram's news environment, offering new insights into engagement dynamics that contribute to attention, polarization, and toxicity amplification.

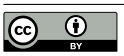
CCS Concepts

• **Information systems** → **World Wide Web**; • **Human-centered computing** → **Social media**.

Keywords

Toxicity, User engagement, Reporting, Bias, Instagram, News

ACM Reference Format:
Sehrish Qummar, Alireza Mohammadinodooshan, and Niklas Carlsson. 2026. Uneven Toxicity Dynamics: A Multi-Dimensional Analysis of Toxicity and Engagement in Instagram News. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3795766.3799776



WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/2026/05
https://doi.org/10.1145/3795766.3799776

1 Introduction

Social media has transformed the news ecosystem by making audience engagement central to content visibility and reach. Rather than acting as passive consumers, users now shape dissemination through likes, comments, shares, and algorithmic feedback loops [19]. This transformation is especially consequential given that social media surpassed television as the primary news source for U.S. adults by 2025 [20]. As a result, news influence is determined not only by editorial choices, but by audience response.

Toxicity as a Driver of Engagement: Toxic and inflammatory language is known to attract attention and boost engagement online, often at the cost of increased polarization and misinformation [2, 26]. Yet, despite growing concern about toxicity in political communication, little is known about how it operates within Instagram's news ecosystem. Understanding whether—and for which publishers and audiences—toxicity is linked to higher engagement on Instagram, therefore, remains a critical open question.

The Unique Context of News on Instagram: Instagram is now the third-largest platform (after Facebook and YouTube) in the United States, serving as a regular news source for 20% of adults [22]. Its emphasis on publisher branding and lightweight interaction (e.g., likes and short comments), rather than explicit political discussion or link sharing, suggests that engagement dynamics may differ from those on Facebook [16], YouTube [10], and X [15]. Yet, despite its growing role, large-scale empirical analyses of news engagement on Instagram remain unexplored, and no prior work examines the outlets' use of toxicity and the corresponding audience response.

Main Contributions: To fill this gap, we present a large-scale, multidimensional analysis of news engagement and toxicity on Instagram, in which we analyze 284,843 posts from 1,026 U.S. news publishers, combining large-scale Instagram engagement data with a multidimensional labeling of outlets by *political bias* and *factual reporting quality*. Our study examines four overarching questions:

- Which kinds of news outlets receive the most (and least) engagement on Instagram? (Section 3)
- Which outlets use the most toxic language in their posts, and how does toxicity vary across reporting quality and political bias? (Sections 4 and 6)
- How does toxicity shape engagement? Do more toxic posts receive disproportionately more interaction, and for whom? (Sections 5 and 6)
- How do audiences respond? Do toxic posts elicit more toxic comments, and does this depend on whether comments agree, disagree, or remain neutral? (Section 7)

Key Example Insights: Across these questions, our analysis yields several central findings: (1) Engagement on Instagram systematically favors *biased* and *lower–reporting-quality* outlets, echoing patterns observed on other platforms but now demonstrated on Instagram. (2) Toxicity is unevenly distributed: low-reporting-quality and politically extreme publishers produce substantially more toxic content. (3) Toxicity functions as an uneven *amplifier*: outlets that already enjoy high engagement—particularly low-reporting and right-leaning outlets—receive disproportionately more engagement on their toxic posts, whereas high-reporting and centrist outlets do not. (4) Toxic posts trigger more toxic comments, with the strongest escalation for right-leaning outlets. This holds even when separating comments by stance (i.e., whether the comments agree, disagree, or are neutral to the post), suggesting toxicity can also operate as in-group reinforcement rather than merely out-group attack.

Implications: These findings deepen our understanding of Instagram's news ecosystem in several ways. For platforms, they highlight which publisher–audience segments are most susceptible to toxicity-driven amplification. For journalists and news organizations, they clarify the engagement incentives that may shape editorial decisions on highly visual platforms. For policymakers and educators, they reveal where toxicity-induced polarization is most likely to emerge, enabling more targeted interventions in media literacy and platform governance.

Roadmap: Section 2 details our methodology. Sections 3–7 present our empirical findings to the above questions. Section 9 reviews related work, and Section 10 concludes.

2 Methodology

2.1 Dataset Construction

Overview: Our dataset was constructed using a structured multi-step pipeline: (1) identifying U.S. news publishers and assigning political-bias and reporting-quality labels; (2) locating and verifying their official Instagram accounts; (3) collecting posts and engagement metrics via CrowdTangle; (4) extracting and preprocessing user comments; (5) computing toxicity scores for posts and comments; and (6) classifying comment stance using a validated zero-shot approach. This pipeline enables systematic comparisons across publisher groups while ensuring transparency and reproducibility.

Identify and Label News Publishers (Step 1): Our dataset is grounded in the sampling of U.S. news publishers rated by Media Bias/Fact Check (MBFC) [13], each labelled along two dimensions: political bias and factual reporting quality. We adopt labels from MBFC, which is a widely used media rating organization in computational social science research (e.g., [4, 31]) that provides broad coverage of U.S. publishers, publicly available and regularly updated ratings, and a transparent methodology [14]. Prior work has shown strong agreement between MBFC and other professional rating systems [11], supporting its external validity.

As of June 2024, we compiled publishers' bias and reporting labels from their corresponding MBFC pages (using descriptive text, assigned tags, and bias-meter indicators). By defining our scope as the universe of MBFC-rated U.S. outlets, we ensure that all analyzed publishers have undergone the same standardized, independent auditing process. However, this selection criteria necessarily excludes smaller outlets not yet indexed by MBFC.

Bias labels were restricted to the five standard, ordered political categories: *Left*, *Left-Center*, *Least Biased*, *Right-Center*, and *Right*. To improve the statistical power, for most analyses, we further aggregate these into three macro-groups: **Left** (*Left*, *Left-Center*), **Center** (*Least Biased*), and **Right** (*Right-Center*, *Right*), and then validated the trends using the five-class groups. Publishers labeled as *Pro-Science*, *Conspiracy-Pseudoscience*, or *Questionable Source* were excluded to focus on the mainstream political spectrum.

MBFC assigns six factual reporting labels (*Very Low*, *Low*, *Mixed*, *Mostly Factual*, *High*, *Very High*). Since some fine-grained categories contain few publishers with active Instagram accounts (e.g., only 11 in *Very Low*), we aggregate them into three ordered macro-groups: **Low Reporting** (*Very Low*, *Low*, *Mixed*), **Mostly Factual**, and **High Reporting** (*High*, *Very High*). This preserves the ordinal structure while ensuring sufficient statistical power. We exclude MBFC's composite *Credibility* score, which conflates bias and reporting.

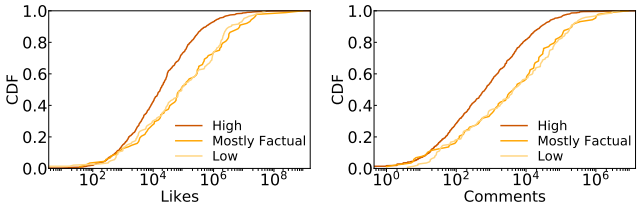
Instagram Account Identification (Step 2): We identified and carefully verified the official Instagram accounts for each publisher using links from their websites or via Instagram's internal search.

Post and Engagement Collection (Step 3): Using the CrowdTangle API [6], we collected all public Instagram posts for a random sample of 1,026 publishers drawn from the total pool of U.S. news outlets identified in Step 1 that maintained a verifiable Instagram presence. This collection took place between Jan. 1 and June 30, 2024, and included each post's engagement metrics. Engagement was retrieved between July 6 and Aug. 13, 2024, ensuring a minimum six-week delay to capture largely converged interactions (over 99% typically accrue within days [23, 28]). Despite CrowdTangle's termination on Aug. 14, 2024, the final dataset spans 1,026 publishers and 284,843 posts, providing sufficient scale for robust statistical analysis.

Comment Extraction (Step 4): To analyze comment content, we retrieved user comments for a stratified random subset of posts. We retained only posts with captions longer than 10 characters to ensure sufficient semantic context, then sampled 1,170 posts with balanced representation across bias–reporting categories. Because CrowdTangle does not support automated comment retrieval, comments were extracted manually, resulting in 182,368 comments.

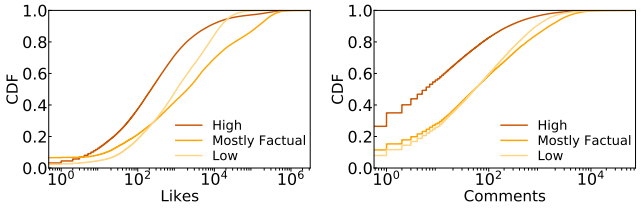
Post and Comment Toxicity Scores (Step 5): After removing posts without text (e.g., image-only posts or empty captions), we computed toxicity scores for all remaining posts and comments using the Google Perspective API, a widely adopted service used by major news organizations such as *The New York Times*, *Le Monde*, and *The Financial Times* [2]. The API is trained on large-scale, multi-platform datasets and outputs continuous scores in $[0, 1]$ that estimate the likelihood that text is rude, disrespectful, or otherwise disruptive; its deterministic and publicly accessible scoring ensures reproducibility. Our analysis focuses primarily on the *toxicity* attribute and its related production attributes: *severe toxicity*, *insult*, *threat*, *profanity*, and *identity attack*.

Stance Classification of Comments (Step 6): We classify each comment's stance toward its associated post as *agree*, *neutral*, or *disagree* using a zero-shot learning approach. After prompt tuning against proprietary state-of-the-art reference models, we labeled post–comment pairs with `gpt-oss-120` (Aug. 2025) [21], selected via a pilot study on 450 stratified samples (50 per bias–reporting



(a) Likes (b) Comments

Figure 1: Per-publisher engagement by Reporting class.



(a) Likes (b) Comments

Figure 2: Per-post engagement by Reporting class.

group). The model achieved 72.9–76.2% agreement with Gemini-3 Pro, GPT-5, and Claude Sonnet 4.5, outperforming other open models (e.g., `Llama-3-70B`). To put these results in context, inter-model agreement among the proprietary models was 79.6–84.7% (mean: 81.71%), indicating a realistic upper bound for automated stance labeling.

To further improve reliability, we also asked the models for confidence scores and retained only labels with confidence above 80%, discarding approximately 18% of samples. This filtering increased agreement to 76.4–82.9% (mean: 79.67%), comparable to confidence-filtered proprietary models. Applying this threshold to the full dataset yielded 145,257 high-confidence post–comment pairs.

2.2 Statistical Testing and Confidence Reporting

Because engagement and toxicity measures are strongly skewed, we use the non-parametric Kruskal–Wallis test to compare group distributions. When significant, we next apply post-hoc Dunn tests for all pairwise comparisons. Throughout the paper, we treat results as statistically significant when $p < 0.01$ and report representative p-values for main findings; e.g., those illustrated with Cumulative Distribution Function (CDFs).

For robustness checks, we provide compact boxplots without p-values; the underlying tests were nearly always consistent with the main results, and we explicitly note the rare exceptions.

3 Baseline Engagement Comparisons

In this section, we benchmark user engagement across two core dimensions used to characterize news publishers: *reporting quality* and *political bias*. Our goal is to establish whether certain types of outlets systematically receive more audience interaction and whether these disparities persist after accounting for structural factors such as publisher follower base and posting volume. Consistent with prior work on platform dynamics, we examine engagement through the primary indicators: likes and comments, supplemented by per-post and normalized per-follower engagement rates.

3.1 Comparing Reporting Classes

We first compare publishers by reporting quality (*High*, *Mostly Factual*, *Low*). As shown in Table 1, the dataset is heavily skewed toward *High*-reporting outlets (n=794; 226 M followers; 193 k posts), with far fewer *Mostly Factual* (88) and *Low* (144) publishers. Despite this skew, engagement intensity differs substantially across groups.

Total Combined Per-Class Engagement: The *High*-reporting group receives the most total engagement, reflecting its larger number of publishers and posts (193 k vs. 59.5 k for *Low* and 32.3 k for *Mostly Factual*). Their posts accumulate 2.1 B likes and 29.4 M comments, compared to 1.31 B likes and 17.3 M comments for *Mostly Factual* and 327 M likes and 20.2 M comments for *Low*.

Per-Publisher Engagement: A very different pattern emerges at the publisher level. As shown in Figure 1, which plots each publisher's median per-post engagement, *High*-reporting outlets exhibit the lowest publisher-level engagement. Their CDFs are shifted far to the left of those for *Mostly Factual* and *Low* outlets, indicating substantially fewer interactions for most *High*-reporting publishers. These differences are statistically supported: Kruskal–Wallis tests show significant overall group effects for both likes ($p = 6.16 \times 10^{-10}$) and comments ($p = 3.25 \times 10^{-11}$). Post-hoc Dunn tests confirm that *High* differs significantly from both *Mostly Factual* and *Low* ($p < 1.74 \times 10^{-5}$ in all cases), while no significant difference is observed between *Mostly Factual* and *Low* ($p = 0.78$ and $p = 0.82$). This confirms that the latter two categories exhibit comparably higher publisher-level engagement, in contrast to the systematically lower engagement of *High*-reporting outlets.

Per-Post Engagement: The engagement differences between reporting groups widen even further at the post level. As shown in Figure 2, the *High*-reporting outlets again receive the lowest engagement, with CDFs lying well to the left of those for the *Mostly Factual* and *Low*. (All pairwise differences are significant at the 0.01 level, including between *Mostly Factual* and *Low*.) Across both likes and comments, posts from the latter two groups attract substantially more interaction across nearly the entire distribution, amplifying the disparities observed at the publisher level.

Accounting for Follower Effects: Because audience size and engagement jointly shape reach, we assess whether follower counts explain the observed gaps. While *Mostly Factual* and *Low* outlets have larger follower bases (Figure 3(a)), normalizing engagement by followers only partially reduces the differences (Figures 3(c) and 3(d)). Both groups still receive more likes and comments per follower than *High*-reporting outlets, indicating that their higher engagement reflects stronger audience responsiveness rather than scale alone. These elevated per-follower rates may, in turn, reinforce follower growth and disparities in audience reach over time.

Controlling for Posting Volumes: Groups that publish more frequently (Figures 3(b)) also tend to achieve higher *per-post* engagement (Figures 2, 3(e), and 3(f)). However, this co-occurrence does not imply that posting frequency drives engagement success. Our publisher-level analyses are based on *median per-post engagement*, which already controls for posting activity. Even under this normalization, *High*-reporting outlets consistently receive the weakest engagement, despite producing the largest number of posts overall (193 k). This result holds across both post-level distributions

Table 1: Summary of engagement metrics across categories.

		Publisher Audience and Activity			Total Engagement		Engagement per Post	
Dimension	Category	Publishers	Followers	Posts	Likes	Comments	Likes	Comments
Reporting	High	794	225,911,871	193,290	2,092,054,544	29,409,651	10,823.4	152.2
	Mostly Factual	88	170,802,032	32,049	1,307,869,203	17,315,706	40,808.4	540.3
	Low	144	103,840,047	59,504	327,333,765	20,243,203	5,501.0	340.2
Bias	Left	406	372,767,561	137,973	3,473,864,823	49,923,687	25,177.9	361.8
	Center	404	43,935,325	86,427	72,659,803	3,555,634	840.7	41.1
	Right	216	83,851,064	60,443	180,732,886	13,489,239	2,990.1	223.2

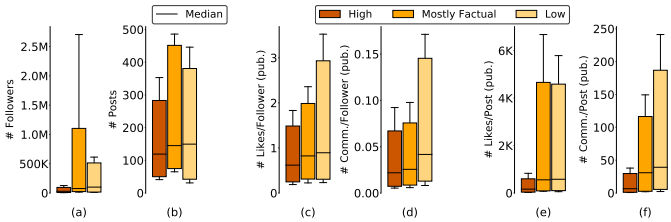
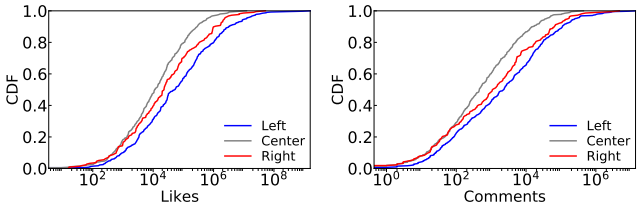
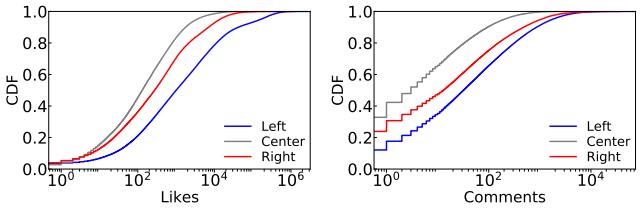


Figure 3: Reporting-based per-follower and per-post engagement (calculated for each publisher): (a) followers/publisher, (b) posts/publisher, (c+d) engagement per follower, and (e+f) average engagement per post (over all the publisher's posts).



(a) Likes **(b) Comments**

Figure 4: Per-publisher engagement by Bias class.



(a) Likes **(b) Comments**

Figure 5: Per-post engagement by Bias class.

(Figure 2) and publisher-level comparisons (Figures 3(e) and 3(f)).[1] Together, these normalizations and observations show that engagement disparities reflect differences in audience response to content rather than differences in posting volume.

3.2 Comparing Bias Classes

We next compare publishers by political bias (*Left*, *Center*, and *Right*). As shown in Table 1, the *Left* category is the largest in publishers (n=406), followers (373 M), and posts (138 k), followed by *Center* (404 publishers, 44 M followers, 86 k posts) and *Right* (216 publishers, 84 M followers, 60 k posts). Despite these size differences, engagement intensity differs systematically across bias groups.

Total Combined Engagement: Given its larger scale, the *Left* group receives the most total engagement: 3.47 B likes and 49.9 M comments across 138 k posts. In comparison, the *Center* group's 86.4 k posts receive 72.7 M likes and 3.6 M comments, and the *Right* group's 60.4 k posts receive 180.7 M likes and 13.5 M comments. While these aggregate differences largely reflect scale, we next examine engagement intensity.

Per-Publisher Engagement: Figure 4 shows that the *Center* outlets consistently have the lowest per-publisher engagement, with CDFs shifted far left of both *Left* and *Right* across both likes and comments (post-hoc Dunn $p < 0.01$ for all cases). Between the biased outlets, *Left*-leaning publishers receive higher engagement than *Right*-leaning publishers (post-hoc Dunn: $p = 0.002$ and $p = 0.031$). Overall, these results indicate that political bias is a strong predictor of engagement: both *Left* and *Right* outlets outperform the *Center* group, with *Left* receiving the most interactions.

Per-Post Engagement: Like in the case of reporting, these patterns become even more pronounced (and differences more significant) at the post level. As shown in Figure 5, *Left*-leaning outlets receive the highest per-post engagement, with CDFs shifted well right of both *Right* and *Center*, and *Right*-leaning outlets again outperform *Center*, which consistently receives the weakest engagement. Overall, politically leaning outlets attract significantly more interaction per post than neutral ones, with *Left* leading.

Accounting for Follower Effects: We next assess whether follower counts explain the observed engagement gaps. Although *Left*-leaning outlets have substantially larger audiences (Figure 6(a)), normalizing engagement by follower count (Figures 6(c) and 6(d)) reduces but does not eliminate the differences. *Left*-leaning outlets still attract significantly more likes per follower than both the *Right* and *Center* groups across most of the distribution. For comments, *Right* exceeds *Left* only in the extreme upper tail, reflecting a small subset of highly active audiences, but *Left* remains ahead for the vast majority of publishers. Overall, these results show that higher engagement among politically leaning outlets is not driven by audience size but by stronger per-user interaction, with users engaging far more intensively than they do with neutral *Center* outlets.

Controlling for Posting Volume: While the groups that publish more frequently (Figure 6(b)) again see the highest *per-post* engagement, these differences do not explain the engagement patterns. As in the reporting analysis, the results are consistent across our publisher-level comparisons (Figure 4), which *median per-post*

[1] Although Table 1 reports a higher mean likes-per-post for *High* than for *Low*, this is driven by a small number of extreme outliers. Across most percentiles, the *Low* CDF lies to the right of the *High* CDF (Figure 2).

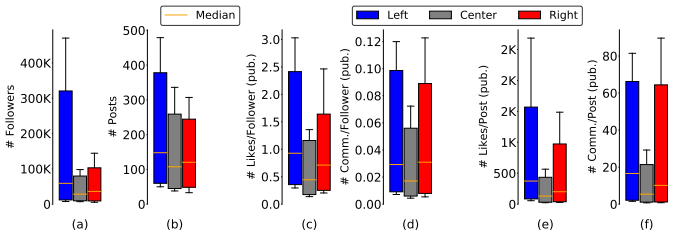


Figure 6: Bias-based per-follower and per-post engagement (calculated for each publisher): (a) followers/publisher, (b) posts/publisher, (c+d) engagement per follower, and (e+f) average engagement per post (over all the publisher's posts).

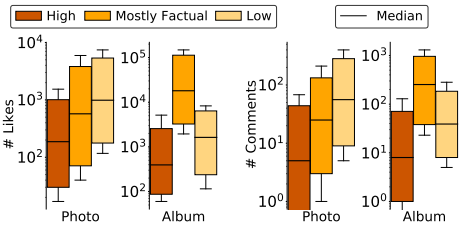


Figure 7: Engagement with different post types (Reporting).

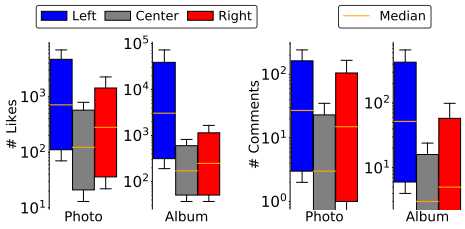


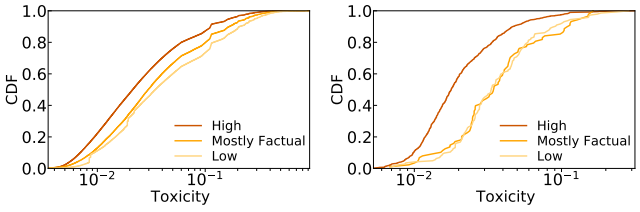
Figure 8: Engagement with different post types (Bias).

engagement already controls for posting frequency, the average per-post (and publisher) values in Figures 6(e) and 6(f), and the per-post distributions in Figure 5. Under each normalization, *Left*-leaning outlets receive by far the highest engagement per post, followed by *Right*, with *Center* outlets consistently trailing. Only in the extreme upper tail of the comments distribution does *Right* briefly surpass *Left*. Together, these results show that engagement disparities are driven by audience responsiveness to politically leaning content rather than by differences in posting volume.

3.3 Generalizations Across Post Types

To test for potential post-type effects, we compare likes and comments across the two dominant post types in our dataset: *photos* and *albums*. Figures 7 and 8 summarize the results.

Across both post types, the engagement hierarchies identified earlier hold. For reporting quality, *Low* and *Mostly Factual* outlets consistently outperform *High*-reporting outlets. For political bias, *Left* and *Right*-leaning outlets receive substantially more engagement than *Center* outlets. Although some gaps narrow slightly across formats, the overall ordering remains unchanged.



(a) Post level **(b) Publisher level**

Figure 9: Toxicity CDFs for each Reporting class.

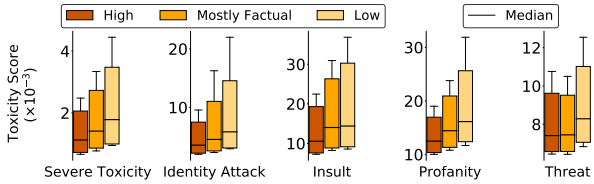


Figure 10: Production attributes for Reporting (post level).

These results show that the engagement patterns documented in Sections 3.1 and 3.2 are not driven by differences in post type: the same patterns persist across Instagram's dominant content formats.

4 Toxicity Usage Comparison

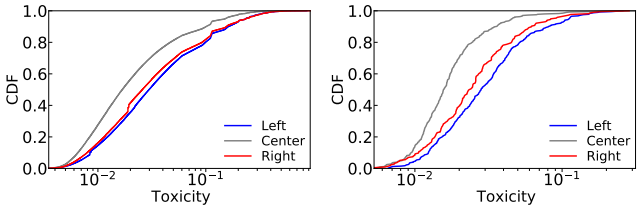
Toxicity offers a distinct perspective on how outlets communicate and how their content may shape online discourse. For example, prior work suggests that toxic language can increase attention while fueling polarization and misinformation [2], making it important to assess whether certain groups rely on such language more heavily. We next investigate how toxic language use varies across publishers with different reporting quality and political orientations, and in Section 5, we investigate how it shapes engagement.

4.1 Reporting-Based Comparison

Post Level: Figure 9(a) shows that *Low*-outlets produce the most toxic content, followed by *Mostly Factual* and then *High*. All pairwise differences are significant ($p < 2.71 \times 10^{-99}$). This pattern aligns with expectations: higher-quality reporting tends to avoid toxic or inflammatory language, whereas lower-quality outlets rely on it more frequently. These differences help contextualize why *Mostly Factual* and *Low* groups receive higher engagement overall.

Publisher Level: At the publisher level (using median toxicity across each outlet's posts; Figure 9(b)), the *Mostly Factual* and *Low* groups show similar and often overlapping toxicity patterns (Dunn: $p = 0.82$), while the *High* group consistently exhibits the lowest toxicity levels ($p = 2.23 \times 10^{-13}$ and $p = 4.10 \times 10^{-21}$).

Generalization Across Production Attributes: The same ordering holds across all toxicity-related attributes: *severe toxicity*, *identity attack*, *insult*, *profanity*, and *threat* (Figure 10). *Low*-reporting outlets consistently are most toxic, followed by *Mostly Factual* and then *High*. This ordering is robust at both post and publisher levels, with only minor attribute-specific reversals between *Low* and *Mostly Factual* at the publisher level (omitted for space).



(a) Post level (b) Publisher level

Figure 11: Toxicity CDFs for each Bias class (3 classes)

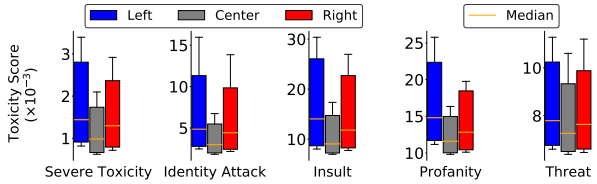
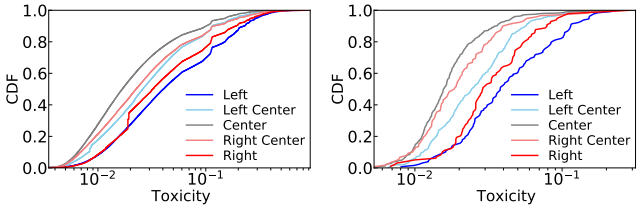


Figure 12: Production attributes for Bias (post level).



(a) Post level (b) Publisher level

Figure 13: Toxicity CDFs for each Bias class (5 classes).

4.2 Bias-Based Comparison

Post Level: At the post level (Figure 11(a)), both *Left* and *Right* publishers exhibit significantly higher toxicity than the more neutral *Center* group. The *Left* and *Right* distributions largely overlap, with *Left* showing slightly higher toxicity across much of the CDF, while *Right* exceeds *Left* in the upper tail. Overall, politically leaning outlets are considerably more toxic than *Center* outlets, with *Left* marginally but yet significantly more so (Dunn: $p = 6.92 \times 10^{-76}$).

Publisher Level: At the publisher level (Figure 11(b)), the distinctions sharpen: *Left* publishers show the highest median toxicity, followed by *Right*, while *Center* remains clearly the least toxic. (Pairwise Dunn: $p = 4.57 \times 10^{-31}, p = 0.0069, p = 3.82 \times 10^{-12}$.) The stronger separation across groups indicates that politically leaning outlets rely more consistently on toxic language than neutral ones.

Generalization Across Production Attributes: Across all five toxicity-related attributes (both post-level in Figure 12 and publisher-level, omitted due to space limit), the *Left* group shows the highest levels of toxic content, the *Right* group intermediate levels, and the *Center* group the lowest.

Further Bias-Based Generalization(s): Across all toxicity measures, a consistent pattern emerges: more politically biased outlets tend to produce more toxic content, with the *Left* group generally exhibiting the highest toxicity levels. To illustrate how this pattern extends beyond the three-class scheme, Figure 13 reports results for a five-class bias model. The groups vary substantially in size—from

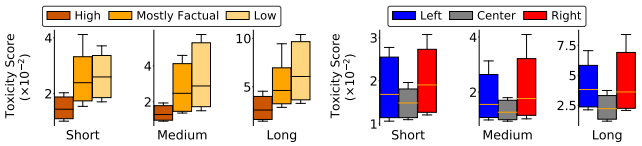


Figure 14: Generalization for different post lengths.

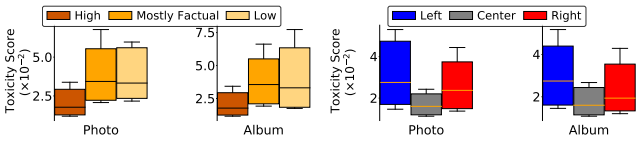


Figure 15: Type-based toxicity comparison based on both Reporting and Bias classes.

45,806 posts and 128 publishers in *Left* to 24,444 posts and 98 publishers in *Right*, with intermediate counts for *Left Center* (92,167 posts, 278 publishers), *Center* (86,427 posts, 404 publishers), and *Right Center* (35,999 posts, 118 publishers)—yet the overall toxicity ordering remains robust. Here, all pairwise Dunn tests were significant (i.e., $p < 0.01$) except for two publisher-level tests: $p = 0.13$ (*Left* vs. *Right*) and $p = 0.049$ (*Center* vs. *Center-Right*).

4.3 Further Robustness Checks

Controlling for Post Length: Since we found that longer posts tend to receive higher toxicity scores, we partitioned all posts into three equally sized groups—short, medium, and long—to control for length effects. However, as seen in Figure 14, the toxicity patterns reported above remain consistent within each length category.

Controlling for Post Type: Toxicity patterns remain consistent across content formats. Whether posts are photos or albums (Figure 15), *Low* and *Mostly Factual* outlets consistently display higher toxicity than *High* in the Reporting dimension, and *Left* and *Right* exceed *Center* in the Bias dimension. These stable gaps indicate that toxicity differences are not driven by post-type composition, but by persistent behavioral tendencies within publisher groups.

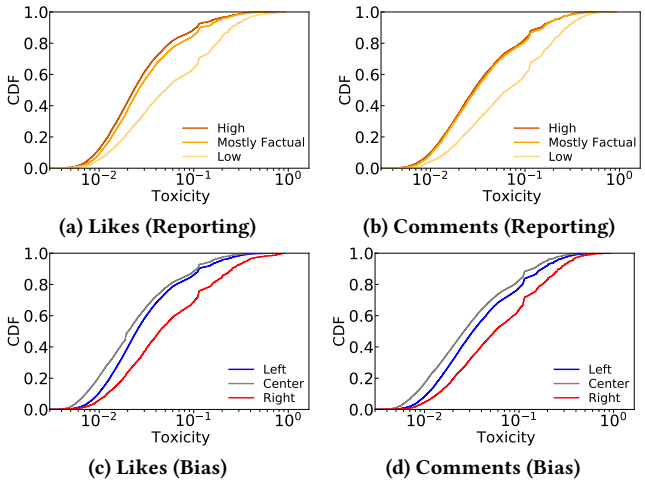
5 Toxicity-Engagement Dynamics

We now examine how audience engagement varies with the toxicity of posts. Section 3 documented baseline engagement differences across Reporting and Bias groups, and Section 4 showed systematic variation in the use of toxic language. Here, we ask whether toxicity itself is associated with disproportionate engagement and whether such effects differ across publisher groups.

5.1 Engagement Share

To quantify audience responsiveness to toxic content, we use an engagement–share measure that tracks, for each group, the fraction of total likes or comments contributed by posts at increasing toxicity levels. This approach highlights whether engagement becomes concentrated on more toxic content within a group. Figure 16 present the resulting CDFs.

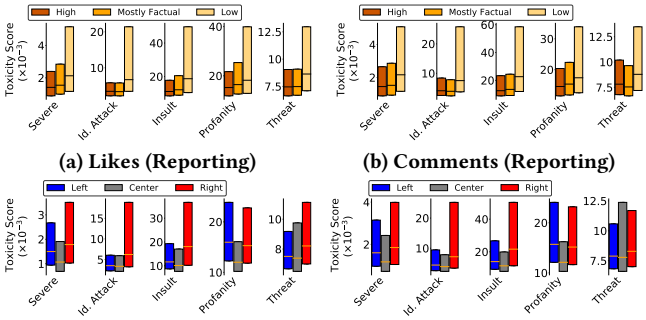
Reporting: Across likes and comments, the *Low*-reporting group derives the largest share of its engagement from toxic posts, followed by *Mostly Factual* and then *High*. Even at moderate toxicity



(a) Likes (Reporting) (b) Comments (Reporting)

(c) Likes (Bias) (d) Comments (Bias)

Figure 16: CDF of the share of engagement with posts of different toxicities, comparing both Reporting and Bias classes.



(a) Likes (Reporting) (b) Comments (Reporting)

(c) Likes (Bias) (d) Comments (Bias)

Figure 17: Production attributes for engagement share.

levels, *Low*-reporting outlets allocate substantially more engagement to toxic content, indicating that their audiences are especially responsive to toxic material.

Bias: Across likes and comments, the *Right* group allocates the largest share of engagement to toxic posts, followed by *Left*, with *Center* showing the least responsiveness. This indicates that politically biased audiences—especially *Right*-leaning ones—respond more strongly to toxic content.

Conclusion: Across both dimensions, groups that already show high baseline engagement (*Low*-reporting and *Right*-leaning outlets) also receive disproportionately high engagement on toxic posts, whereas *High*-reporting and *Center* groups receive comparatively little. This suggests that *audience responsiveness*, rather than production alone, drives the amplification of toxic content. This dynamic is further augmented by a notable asymmetry between *Left* and *Right* publishers. While *Left*-leaning outlets lead in baseline engagement (Section 3.2), the *Right* group takes the lead in engagement share here. This shift indicates that the *Right*-leaning audience is significantly more responsive to toxic content, whereas the *Left*-leaning audience's high engagement is more evenly distributed across toxic and non-toxic content.

5.2 Production Attributes

Having shown that engagement concentrates on posts with higher overall toxicity, we now test whether this pattern holds across the individual toxicity attributes provided by the Perspective API: *severe toxicity*, *identity attack*, *insult*, *profanity*, and *threat*. These attribute-level analyses (Figure 17) primarily validate the overall trends but also reveal several notable deviations.

Reporting: Across nearly all production attributes, the results closely mirror those for overall toxicity: the *Low*-reporting group consistently captures the largest engagement share on more toxic posts, while *High* and *Mostly Factual* receive much smaller—and often overlapping—shares. This pattern is especially clear for *severe toxicity*, *insult*, and *profanity*, where the ordering *Low* > *Mostly Factual* > *High* appears with little ambiguity.

Some attributes show modest deviations. For *identity attack*, the *High* and *Mostly Factual* groups overlap for most of the range, with *High* occasionally surpassing *Mostly Factual* at lower toxicity levels. For *threat*, the two groups are nearly indistinguishable, suggesting similar audience responses among higher-quality outlets. These small departures do not alter the broader pattern but indicate that the *High*–*Mostly Factual* ordering is attribute-dependent.

Comment-based engagement follows the same structure: *Low* remains clearly ahead, while *High* and *Mostly Factual* mostly overlap. Stronger separation appears for *profanity*, and for *threat*, where *Mostly Factual* tends to fall slightly behind *High*.

Bias: The Bias dimension shows clearer differentiation across attributes. For *severe toxicity*, *identity attack*, and *insult*, the *Right* group consistently captures the largest engagement shares, with *Center* lowest—reinforcing the overall toxicity pattern.

Some attributes, however, deviate from this ordering. For *profanity*, the *Right* group drops toward the lower end of the distribution, indicating reduced responsiveness to this form of toxicity. For *threat*, group differences narrow considerably, with substantial overlap across the distribution. Comment-based engagement follows the same general structure: the leaning groups typically outrank *Center*, though *threat* again produces the greatest convergence.

Conclusion: Across attributes, audiences of *Low*-reporting and *Right*-leaning outlets consistently show the strongest engagement with toxic content. Although a few attributes—particularly *profanity* and *threat*—introduce minor deviations, these do not alter the overall pattern: toxic content is amplified most among *Low*-quality and *Right*-leaning publishers.

5.3 Robustness: Bootstrapping

To test whether engagement–share patterns are driven by posting volume rather than audience responsiveness, we apply a bootstrapping procedure that equalizes posting activity across publishers by repeatedly sampling the same number of posts per outlet and recomputing engagement shares.

The results closely track the post-level findings (Figure 18): *Low*-reporting and *Right*-leaning outlets continue to receive the largest engagement share on toxic posts, while *High*-reporting and *Center* outlets remain least responsive. Although differences between *High* and *Mostly Factual* narrow slightly after equalization, the overall ordering is unchanged. This robustness to the sampling method

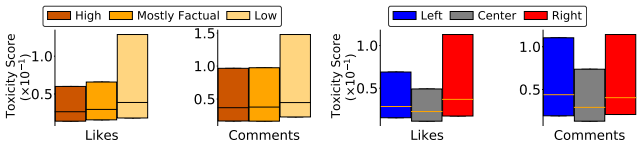


Figure 18: Robustness check using bootstrapping.

reinforces that toxicity-driven engagement disparities reflect audience behavior rather than differences in posting volume.

6 Two-Dimensional Analysis: Reporting × Bias

The preceding sections examined Reporting and Bias dimensions independently, revealing clear patterns in use of toxic language (Section 4) and the relationship between toxicity and engagement (Section 5). We next study how these patterns behave when *both* dimensions are considered jointly. Specifically, we analyze toxicity and toxicity-related engagement across a 3×3 classification matrix defined by Reporting quality (*High*, *Mostly Factual*, *Low*) and Bias (*Left*, *Center*, *Right*), producing nine distinct publisher groups. Across all analyses, we find that the major one-dimensional results remain robust in the two-dimensional setting. However, the joint analysis also reveals several *cross-dimensional interactions* that are not visible when examining Reporting or Bias alone. We highlight these confirmations and new insights in the two subsections below.

6.1 Toxicity Usage Comparison

Figure 19 shows the per-post toxicity usage distributions for Reporting groups within each Bias category.

Confirmations: The two-dimensional results strongly validate the findings of Section 4: (i) *Low*-reporting outlets are the most toxic across all Bias categories, and (ii) *High*-reporting outlets remain the least toxic. Thus, the previously observed monotonic relationship between Reporting quality and toxicity holds even after conditioning on bias. Similarly, our two-dimensional analysis conditioning on Reporting (not shown) confirms that (iii) Bias effects persist, with *Left*- and *Right*-leaning groups being more toxic than the *Center*.

New Insights: Several deviations appear only when both dimensions are considered jointly:

- ***Center*-Bias Anomaly.** In the *Center* bias group, the *Mostly Factual* reporting class becomes less toxic than *High* in the upper percentiles. This inversion does not arise in either one-dimensional analysis.
- **Wider Publisher-Level Separation on the *Right*.** Within the *Right* bias group, toxicity differences between *Low*, *Mostly Factual*, and *High* reporting outlets become markedly sharper, indicating a stronger link between reporting quality and toxicity for *Right*-leaning publishers.

A closer look at individual toxicity attributes (analyses omitted for space) sheds light on the *Center*-bias anomaly: while most attributes follow the aggregate toxicity ordering, the attributes *identity attack* and *threat* display reversals within the *Center* group. These attribute-specific shifts do not appear in Section 4, but explain why the joint analysis uncovers a unique inversion for the *Center* group.

Summary: The monotonic reporting-quality gradient in toxicity is stable, but new relationships emerge within the *Center* groups.

Table 2: Summary statistics of comments and their relative stance across sampled posts (full and > 80% confidence).

Reporting	Bias	# Comments		Agree (%)	Neutral (%)	Disagree (%)
		Full	> 80%			
High	Left	24015	18670	56.1	20.5	23.3
	Center	7613	6224	54.1	19.5	26.3
	Right	5871	4596	54.7	20.5	24.6
Mostly Factual	Left	46084	35936	57.0	19.8	23.1
	Center	11479	9270	45.2	31.5	23.2
	Right	8632	6857	51.5	16.1	32.2
Low	Left	27286	22384	51.2	26.0	22.7
	Center	9677	7956	60.3	27.0	12.5
	Right	41711	33364	52.9	19.7	27.3

6.2 How Toxicity Shapes Engagement

Figure 20 reports engagement share as a function of toxicity for each Reporting group within each Bias class.

Confirmations: The two-dimensional results reinforce the main conclusions of Section 5: (i) *Low*-reporting outlets tend to receive a larger engagement share of on toxic posts. (ii) *Right*-leaning audiences are generally more responsive to toxic content.

New Insights: The interaction between Bias and Reporting reveals patterns unavailable in the aggregate analyses:

- ***Right*-Bias Reversal.** In the *Right* bias group, *Mostly Factual* receives a larger share of engagement on toxic posts than *Low* for most of the toxicity range (up to ~75%), with *Low* dominating only in the extreme tail. This reversal does not appear in either the Reporting-only or Bias-only analyses.
- ***High* Reporting Overtakes *Mostly Factual* in *Center* Bias.** For both likes and comments, *High* occasionally attracts more engagement on toxic posts than *Mostly Factual* within the *Center* bias class.

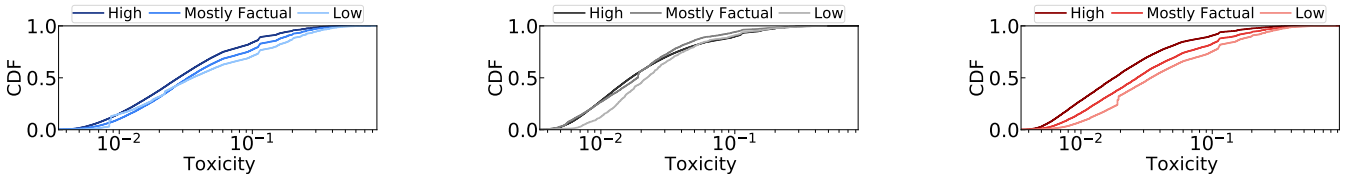
Attribute-level analyses (omitted for space) help explain the *Right*-bias reversal: within the *Right* cluster, different forms of toxicity elicit distinct audience responses. Attributes such as *identity attack* tend to favor *Mostly Factual* outlets, whereas *insult* favors *Low*. These asymmetries do not appear in the one-dimensional analyses of Section 5, but they surface once Reporting and Bias are examined jointly, producing the observed reversal in engagement shares.

Summary: Although the overall relationship between toxicity and engagement is stable across dimensions, the two-dimensional analysis reveals that audience responsiveness to toxic content is *politically dependent* and interacts meaningfully with Reporting quality. *Right*-leaning audiences in particular display complex patterns in how they reward toxicity, with different reporting classes leading at different toxicity levels.

7 Comment-Based Analysis

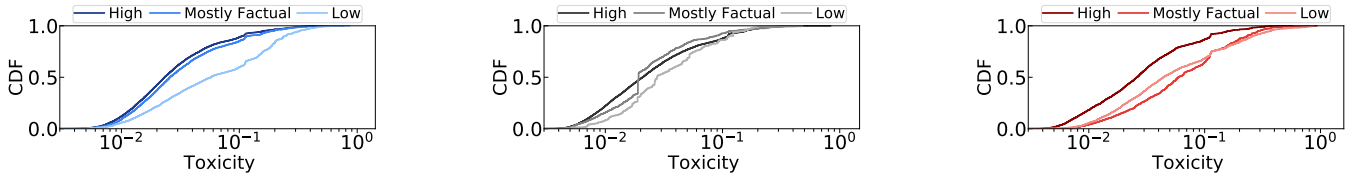
To examine audience response to toxic content, we analyze comments on 1,170 manually sampled posts drawn evenly from all nine Reporting × Bias groups (Section 2). Each comment is labeled for toxicity and stance toward the post (*agree*, *disagree*, *neutral*). Table 2 reports the number of analyzed comments and their stance distribution (restricted to labels with ≥80% confidence).

Comment Toxicity Mirrors and Sharpens Two-Dimensional Toxicity Patterns: Comment toxicity closely follows, and often

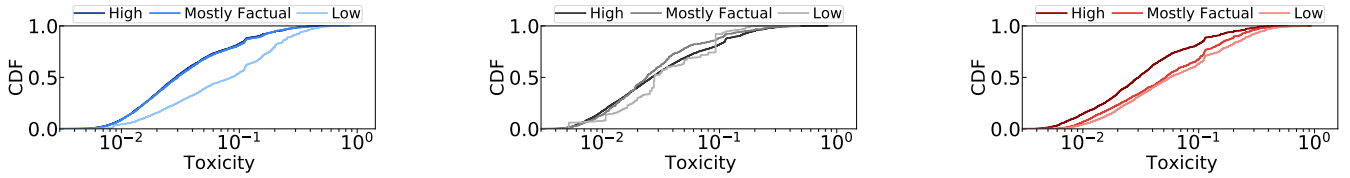


(a) Left (b) Center (c) Right

Figure 19: Toxicity per-post as compared for each Reporting class, conditioned on Bias class. (Per-publisher results are similar.)



(a) Left (likes) (b) Center (likes) (c) Right (likes)



(d) Left (comments) (e) Center (comments) (f) Right (comments)

Figure 20: CDFs of engagement share associated with each toxicity level, comparing Reporting class, conditioned on Bias.

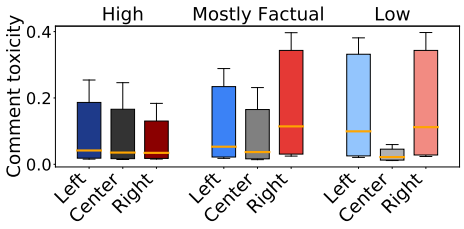


Figure 21: Two-dimensional toxicity overview of comments.

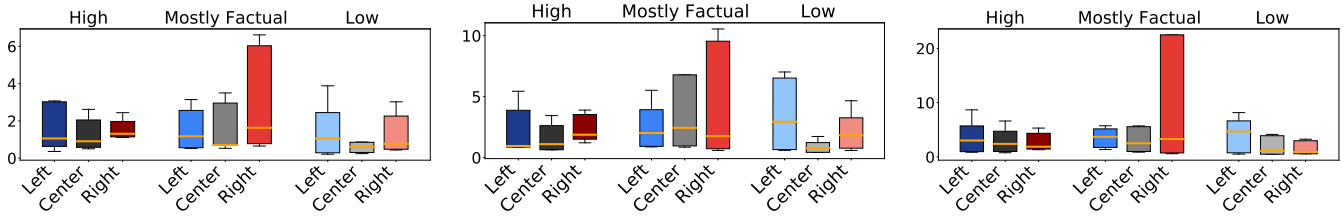
amplifies, the post-level patterns from Section 6. For example, Figure 21 shows that posts from the *Left*- and *Right*-leaning publishers attract more toxic replies than *Center* outlets, and that this separation goes from non-significant for *High*-reporting outlets to highly pronounced for *Low*-reporting publishers, suggesting that audience reactions to these posts may be even more polarized than the content itself. The suggested polarization around politically-biased *Low*-quality outlets is further supported (as seen in Table 2) by *disagreement* being far more common in the *Low*-reporting class for the *Left* and *Right* groups (22–27%) than for *Center* (12.5%).

Toxicity Patterns Persists Across Comment Stances: As expected, disagreement comments are generally the most toxic. Yet, across all stances (*agreeing*, *neutral*, and *disagreeing*; Figures 22(a–c)), the same *Left*/*Right* vs. *Center* toxicity gap appears, with the differences again going from non-significant for the *High*-reporting outlets to being substantial (and significant) for the *Low*-reporting class. Notably, agreeing comments show some of the strongest *Left*/*Right* separation, suggesting reinforcement dynamics within partisan audiences. Disagreement toxicity also exhibits asymmetries: it is higher in *Left*-bias groups for *High* and *Low* reporting, and higher in *Right*-bias groups for *Mostly Factual*, revealing uneven resistance patterns across the spectrum.

Toxic Agreement as In-Group Amplification: Agreement comments constitute the majority of responses (45–60%; Table 2), yet they can still be highly toxic, especially for the *Low*-reporting class. This suggests that toxicity often functions as *in-group signaling*, with supportive commenters amplifying rather than moderating the negativity of the posts they endorse. Such patterns suggest a mechanism of intra-group rhetorical escalation rather than toxicity being driven solely by out-group hostility. Having said that, we have seen that the *Right*-leaning outlets see the biggest escalation in toxicity, as seen by a higher ratio between the toxicity of the comments to the toxicity of the original posts (Figure 23).

Engagement-Share Patterns Validate the Post-Level Toxicity-Engagement Link: Figure 24 shows, across all stance categories, how much of each group's comment activity is directed at toxic posts. The patterns closely mirror the post-level results: the *Low*-reporting group concentrates a larger share of its comments on toxic posts, and posts from *Right*-leaning outlets attract disproportionately many comments on toxic content regardless of stance. These findings reinforce the main conclusion from Section 5: users interacting with posts from *Low*-quality or *Right*-biased outlets respond most intensely to toxic material.

Summary: Overall, comment-level data validate the broad two-dimensional findings of Sections 6.1 and 6.2, while revealing new mechanisms: (1) toxicity is amplified by audience response, especially at ideological extremes; (2) agreeing comments can be as toxic



(a) Agree (b) Neutral (c) Disagree

Figure 22: Stance-based comparison of the toxicity level of the comments associated with posts by publishers with nine different classes (3 Reporting × 3 Bias classes).

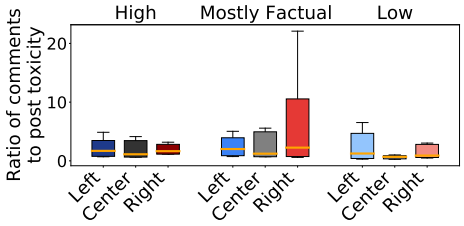
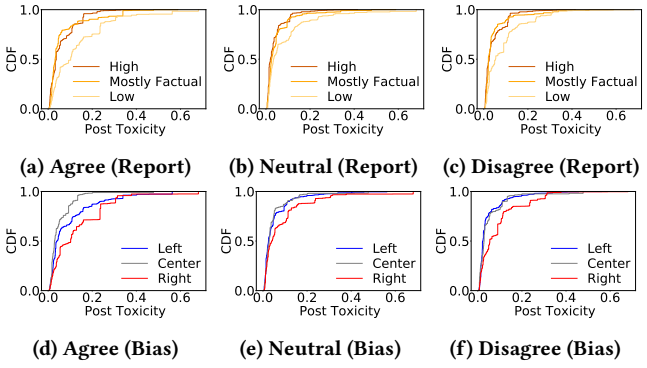


Figure 23: Ratio of comment-to-post toxicity.



(a) Agree (Report) (b) Neutral (Report) (c) Disagree (Report)

(d) Agree (Bias) (e) Neutral (Bias) (f) Disagree (Bias)

Figure 24: Engagement share of toxic comments for (a+b+c) Reporting and (d+e+f) Bias, split across different stances.

as disagreeing ones; and (3) *Low*-reporting and *Right*-leaning communities show the strongest toxicity-engagement coupling. These insights highlight the user behavior's role, not just publisher behavior, in shaping the toxicity dynamics of online news ecosystems.

8 Discusion of Limitations and Future Work

Our study has several limitations that provide opportunities for future research. First, our interpretations of in-group signaling and rhetorical escalation (Section 7) are inferred from comments' stance and toxicity patterns. Because we do not have data on individual user identities or their long-term interaction histories, these mechanisms can not be formally verified. Future research could employ longitudinal user-level tracking or interviews to better understand the psychological drivers of toxic engagement.

Second, regarding our comment-based analysis (Section 7), our decision to exclude low-confidence (< 80% confidence) samples was a deliberate tradeoff to increase the internal validity of our stance-based findings. We acknowledge that ambiguous, mixed, or sarcastic comments may be significant in toxicity dynamics and may represent a unique form of engagement. However, by focusing on high-confidence "clear" cases, we establish a conservative baseline for how toxicity operates across distinct ideological responses. Future research could specifically target these ambiguous interactions to determine if they serve as "bridge" content or further fuel conversational escalation.

Third, our study is based on a correlational design, which does not support definitive causal claims. Although we controlled for various available variables, future research employing experimental frameworks or instrumental variable methods can formally verify the causal mechanisms.

Fourth, while we synchronized our collection of MBFC ratings (June 2024) with our observation window (Jan.–June 2024) to ensure temporal relevance, a potential concern is whether these labels remain valid over longer periods. However, research indicates that domain-level assessments are institutionally stable and robust for longitudinal analysis because they reflect deep-seated organizational characteristics rather than transient content shifts [12]. The high correspondence observed across different expert rating platforms [11] further suggests that the multi-dimensional dynamics identified in this study are rooted in stable institutional behaviors and long-term journalistic standards.

Fifth, our toxicity analysis relies on Google's Perspective API. As a proprietary and commercial tool, it may possess inherent biases or limitations in its underlying models that are not fully transparent. Nevertheless, our approach is supported by significant validation in the literature; studies by Welbl et al. [30] and Wang et al. [29] demonstrate that Perspective API scores correlate strongly with human evaluations of toxicity. Furthermore, Cima et al. [5] report a high correlation between the Perspective API and Detoxify [9], the other widely adopted open-source framework for toxicity detection. These cross-method correlations suggest that our toxicity metrics are consistent with broader academic standards.

Finally, in our joint analysis of Reporting quality and Bias (Section 6), we acknowledge an uneven distribution of publishers. While six of the nine categories contain a substantial number of publishers ($N > 41$), three specific groups—*Low-Center*, *Mostly Factual-Center*, and *Mostly Factual-Right*—consist of 3, 6, and 23 publishers, respectively. However, it is important to note that even within these segments, the dataset includes a high volume of observations, with 728, 2,726, and 7,192 posts respectively, along with the thousands of associated user comments detailed in Table 2. This substantial volume of underlying data supports the statistical robustness of the patterns and interactions observed in Section 6.

9 Related Work

News Engagement on Social Media: A growing body of research has examined how users engage with political news across platforms such as YouTube [10], Facebook [7, 15], and X/Twitter [3, 15, 17, 33]. These studies consistently find that engagement varies systematically with publisher characteristics, including political bias and source reliability. For example, similar to our finding, but on X, it has been shown that unreliable publishers receive significantly more interactions than their reliable counterpart [15]. For Facebook specially, it has been shown that user engagement patterns differ between public and non-public interaction spaces [16]. Beyond single platforms, cross-platform analyses show that partisan and low-quality news tend to exhibit similar engagement advantages across environments [18]. Yet, despite Instagram's prominence as a news venue, empirical evidence on Instagram news is scarce. The largest relevant study focuses on sentiment and topical variation across just 53 outlets and does not examine toxicity or audience response [1].

Toxicity in Political News and Online Discourse: A parallel line of work analyzes toxic language in political communication, particularly on Twitter/X, using large-scale models such as the Google Perspective API [24, 25, 27]. These studies identify linguistic and behavioral patterns associated with toxic users and conversations, and examine toxicity around major news outlets [27]. Related work on Reddit employs neural toxicity classifiers (e.g., Detoxify) to study toxic content and user behavior [32].

Toxicity and User Response: A smaller but growing body of research jointly examines toxicity and engagement. Findings are mixed: while Falkenberg et al. [8] report that toxicity correlates with reduced engagement across nine countries, other work shows that toxic content can stimulate interaction. For example, Beknazar et al. [2] measure engagement with toxic posts across multiple platforms in a field experiment, and Salehabadi et al. [26] demonstrate that toxic replies on Twitter/X propagate further toxicity, reinforcing hostile conversational dynamics.

Our Contribution: This paper distinguishes itself from prior work in three ways: (1) we provide the first large-scale analysis of news engagement and toxicity on Instagram, a major platform that has been largely overlooked; (2) we introduce a multi-dimensional framework that jointly considers political bias, factual reporting quality, and toxicity to understand their combined effect on user interaction; and (3) we move beyond post metrics by analyzing the toxicity and stance of user comments, revealing how toxic news content shapes toxic audience response.

10 Conclusion

This paper presents the first large-scale analysis of news engagement, toxicity, and audience response on Instagram. Analyzing 284,843 posts from 1,026 U.S. news publishers, we show that engagement systematically favors politically extreme and lower–reporting-quality outlets, demonstrating that engagement hierarchies observed on other platforms extend to Instagram's news environment.

We further find that toxicity is strongly associated with amplified and uneven engagement. Toxic language is disproportionately produced by low-reporting and ideologically extreme publishers, particularly on the right, and attracts a disproportionate share of engagement for outlets that already receive high baseline attention. In contrast, high-reporting and centrist outlets do not experience comparable engagement gains from toxic content, indicating that toxicity interacts with audience predispositions rather than uniformly increasing attention.

Audience responses reinforce this pattern. Toxic posts elicit significantly more toxic comments, with the strongest escalation among right-leaning publishers. This effect persists across reporting quality and comment stance, suggesting that toxicity often serves as in-group reinforcement rather than out-group hostility. Together, these findings reveal a feedback loop in which toxicity, audience alignment, and engagement reinforce one another.

Overall, our findings highlight uneven engagement and polarization dynamics associated with toxic news content on Instagram, with important implications for platform governance, journalism, and media literacy.

Acknowledgments

The authors express their thanks to the anonymous reviewers for their insightful comments that helped improve the paper. This work was funded by Excellence Center at Linköping-Lund on Information Technology (ELLIIT, project C05 GPAI), Graduate School in Computer Science (CUGS), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Kholoud Khalil Aldous, Jisun An, and Bernard J. Jansen. 2023. What really matters?: characterising and predicting user engagement of news postings using multiple platforms, sentiments and topics. *Behaviour & Information Technology* 42, 5 (2023), 545–568. doi:10.1080/0144929X.2022.2030798

[2] George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. *Toxic content and user engagement on social media: Evidence from a field experiment.* Technical Report 11644. CESifo Working Paper.

[3] Andrea K. Bellovary, Nathaniel A. Young, and Amit Goldenberg. 2021. Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly. *Affective Science* 2, 4 (2021), 391–396. doi:10.1007/s42761-021-00046-w

[4] Tejasvi Chebrolu, Rohan Modepalle, N Harsha Vardhan, Ponnurangam Kumaraguru, and Ashwin Rajadesingan. 2025. Framing the Fray: Conflict Framing in Indian Election News Coverage. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25).* Association for Computing Machinery, New York, NY, USA, 294–305. doi:10.1145/3717867.3717900

[5] Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2024. The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference* (Stuttgart, Germany) *(Websci Companion '24).* Association for Computing Machinery, New York, NY, USA, 85–93. doi:10.1145/3630744.3663608

[6] CrowdTangle Team. 2021. CrowdTangle. https://www.crowdtangle.com/

[7] Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. Understanding engagement with U.S. (mis)information news sources on Facebook. In *Proceedings of the 21st ACM Internet Measurement Conference* (Virtual Event) *(IMC '21).* Association for Computing Machinery, New York, NY, USA, 444–463. doi:10.1145/3487552.3487859

[8] Max Falkenberg, Fabiana Zollo, Walter Quattrociocchi, Jürgen Pfeffer, and Andrea Baronchelli. 2024. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nature Communications* 15, 1 (2024), 9560.

[9] Laura Hanu and Unitary Team. 2020. Detoxify. https://github.com/unitaryai/detoxify GitHub repository.

[10] Xuejin Jiang, Liming Liu, Biying Wu-Ouyang, Long Chen, and Han Lin. 2024. Which storytelling people prefer? Mapping news topic and news engagement in social media. *Computers in Human Behavior* 158 (2024), 108248. doi:10.1016/j.chb.2024.108248

[11] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2023. High level of correspondence

across different news domain quality rating sets. *PNAS Nexus* 2, 9 (09 2023), pgad286. doi:10.1093/pnasnexus/pgad286

[12] Jula Lühring, Hannah Metzler, Ruggero Lazzaroni, Apeksha Shetty, and Jana Lasser. 2025. Best practices for source-based research on misinformation and news trustworthiness using NewsGuard. *Journal of Quantitative Description: Digital Media* 5 (2025). doi:10.51685/jqd.2025.003

[13] Media Bias/Fact Check. [n. d.]. Media Bias/Fact Check. https://mediabiasfactcheck.com Accessed: 2024-06-10.

[14] Media Bias/Fact Check . 2025. Media Bias/Fact Check Methodology. https://mediabiasfactcheck.com/methodology

[15] Alireza Mohammadinodooshan and Niklas Carlsson. 2024. Understanding Engagement Dynamics with (Un)Reliable News Publishers on Twitter. In *Social Networks Analysis and Mining: 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III* (Rende, Italy). Springer-Verlag, Berlin, Heidelberg, 36–47. doi:10.1007/978-3-031-78548-1_4

[16] Alireza Mohammadinodooshan and Niklas Carlsson. 2025. Public versus Less-Public News Engagement on Facebook: Patterns Across Bias and Reliability. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25)*. Association for Computing Machinery, New York, NY, USA, 437–448. doi:10.1145/3717867.3717912

[17] Alireza Mohammadinodooshan and Niklas Carlsson. 2026. Sentiment-Driven Differential Engagement: Hyperpartisan Vs. Non-hyperpartisan Users on X. In *Social Networks Analysis and Mining*, Aijun An, Alfredo Cuzzocrea, and Hongxin Hu (Eds.). Springer Nature Switzerland, Cham, 101–118.

[18] Mohsen Mosleh, Jennifer Allen, and David G. Rand. 2025. Divergent patterns of engagement with partisan and low-quality news across seven social media platforms. *Proceedings of the National Academy of Sciences* 122, 44 (2025), e2425739122. doi:10.1073/pnas.2425739122

[19] Subhayan Mukerjee, Tian Yang, and Yilang Peng. 2023. Metrics in action: how social media metrics shape news production on Facebook. *Journal of Communication* 73, 3 (04 2023), 260–272. doi:10.1093/joc/jqad012

[20] Nic Newman, Amy Ross Arguedas, Craig T. Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. Digital News Report 2025. *Reuters Institute* (2025). https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025

[21] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 [cs.CL] https://arxiv.org/abs/2508.10925

[22] Pew Research Center. 2025. Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

[23] Jürgen Pfeffer, Daniel Matter, and Anahit Sargsyan. 2023. The Half-Life of a Tweet. *Proceedings of the International AAAI Conference on Web and Social Media* 17, 1 (Jun. 2023), 1163–1167. doi:10.1609/icwsm.v17i1.22228

[24] Hina Qayyum, Muhammad Ikram, Benjamin Zi Hao Zhao, Ian D. Wood, Nicolas Kourtellis, and Mohamed Ali Kaafar. 2023. Exploring the Distinctive Tweeting Patterns of Toxic Twitter Users. In *2023 IEEE International Conference on Big Data (BigData)*. 3624–3633. doi:10.1109/BigData59044.2023.10386402

[25] Hina Qayyum, Benjamin Zi Hao Zhao, Ian Wood, Muhammad Ikram, Nicolas Kourtellis, and Mohamad Ali Kaafar. 2023. A longitudinal study of the top 1% toxic Twitter profiles. In *Proceedings of the 15th ACM Web Science Conference 2023* (Austin, TX, USA) *(WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 292–303. doi:10.1145/3578503.3583619

[26] Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User Engagement and the Toxicity of Tweets. arXiv:2211.03856 [cs.SI] https://arxiv.org/abs/2211.03856

[27] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1086–1097. doi:10.1145/3442381.3449861

[28] Elin Thorgren, Alireza Mohammadinodooshan, and Niklas Carlsson. 2024. Temporal Dynamics of User Engagement on Instagram: A Comparative Analysis of Album, Photo, and Video Interactions. In *Proceedings of the 16th ACM Web Science Conference* (Stuttgart, Germany) *(WEBSCI '24)*. Association for Computing Machinery, New York, NY, USA, 224–234. doi:10.1145/3614419.3644029

[29] Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. 2022. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2595, 14 pages.

[30] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 2447–2469. doi:10.18653/v1/2021.findings-emnlp.210

[31] Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25)*. Association for Computing Machinery, New York, NY, USA, 127–137. doi:10.1145/3717867.3717903

[32] Niloofar Yousefi, Nahiyan Bin Noor, Billy Spann, and Nitin Agarwal. 2024. Examining Toxicity's Impact on Reddit Conversations. In *Complex Networks & Their Applications XII*, Hocine Cherifi, Luis M. Rocha, Chantal Cherifi, and Murat Donduran (Eds.). Springer Nature Switzerland, Cham, 401–411.

[33] Simon Zollo, Matteo Cinelli, Gabriele Etta, Roy Cerqueti, and Walter Quattrociocchi. 2025. Inference of social media opinion trends in 2022 Italian elections. *Expert Syst. Appl.* 269, C (April 2025), 12 pages. doi:10.1016/j.eswa.2024.126377