Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Disproportionate Voices: Participation Inequality and Hostile Engagement in News Comments

Sangbeom Kim[1], **Senhye Noh**[2]
[1]Seoul National University
[2]University of California, Los Angeles
tkdwlwhs@snu.ac.kr, shnoh@g.ucla.edu

Abstract

Digital platforms were expected to foster broad participation in public discourse, yet online engagement remains highly unequal and underexplored. We ask whether *who talks the most* also shapes *the tone that becomes visible* in news comment sections. Using 260 million comments from 6.2 million users over 13 years on *Naver News*, we measure participation inequality with the Gini and Palma indexes and estimate hostility levels with a *KC-Electra* model, which outperformed other Korean pre-trained transformers in multi-label classification tasks. We test two hypotheses: (H1) a within-section, over-time association between participation concentration and hostile tone; and (H2) a decomposition of hostility changes into *composition* (who talks how much) versus *behavior* (how those same groups talk). We find that higher concentration is associated with higher hostile tone, and that changes in hostility are driven mainly by behavioral intensification among active users, rather than compositional shifts. These results *quantify* the relationship suggested by exploratory patterns and point to an amplification risk: when a few users dominate, visible discourse can skew sharper, potentially discouraging casual participation and reinforcing concentration.

Code and Data —
https://github.com/SangbeomKim7/Disproportionate-Voices-ICWSM26

Introduction

Digital platforms were once expected to foster broad and equitable participation in public discourse (Papacharissi 2004). However, growing evidence suggests that online engagement remains highly unequal, with a small fraction of users dominating digital conversations, potentially skewing public discourse (e.g., Van Mierlo 2014; Gasparini et al. 2020; Carron-Arthur, Cunningham, and Griffiths 2014; Baqir et al. 2023; Antelmi, Malandrino, and Scarano 2019). The '90-9-1' principle, although not rigorously tested, suggests a significant disparity in online participation, where 90% of users ('lurkers') primarily observe without participating, 9% ('contributors') engage occasionally, and a mere 1% ('superusers') generate the majority of online content (Nielsen 2006).

This study examines the digital participation divide and its relationship with hostile engagement in online news discussions. Using a 13-year dataset from *Naver News*, South Korea's largest news aggregation platform, we analyze 260 million comments from 6.2 million users to assess the participation inequality between frequent and infrequent commenters in news comment sections and its connection with content hostility. We employ the Gini and Palma indices to quantify participation disparities and apply a transformer-based deep learning model (*KC-Electra*), fine-tuned for multi-label classification, to classify comment hostility levels.

Building on descriptive evidence of highly concentrated participation and our exploratory hostile-content classification, we ask whether *who talks the most* also shapes *the tone that becomes visible*. To make this link concrete, we align measurement on the same activity groups—**Palma** for inequality (Top 10% vs. Bottom 40%) and a **group-conditional** hostility index that tracks the Top 1%, the Top 1–10% (excluding Top 1%), and the Bottom 40% within each section–month (with Bottom 40% used as a *lower-base* group). This alignment keeps the focus on the parts of the conversation that are most likely to shape what readers actually see when participation is concentrated.

The findings reveal a highly unequal participation structure, with a small number of frequent users contributing disproportionately to news comment sections. This participation divide is particularly pronounced in political news domains and in more widely read news stories. Moreover, frequent commenters are significantly more likely to post hostile content, including both uncivil and hateful content, suggesting that online discourse is shaped disproportionately by a highly active and often hostile subset of users. Consistent with this picture, we show—via **H1**—that higher participation concentration is *associated* with higher hostile tone within sections over time, and—via **H2**—that changes in hostility are driven primarily by *behavioral* intensification among active users rather than by shifts in who is present.

This study makes a novel contribution by systematically linking participation inequality with multiple forms of hostile engagement at scale, using user-level trace data over 13 years. Methodologically, we align inequality and hostility on identical activity groups and move from descriptive hints to **testable statements** with transparent measurement, quantifying a robust association that prior work often treated

separately. Substantively, we reframe participation inequality as a *quality-of-discourse* issue: when a few users dominate, the visible conversation can skew sharper, potentially discouraging casual participation and reinforcing concentration. While our design is non-causal and subject to potential confounding and measurement error, the results are *consistent* with an amplification risk that platforms and editors should monitor, and they lay a baseline for future causal designs and policy experiments.

Digital Divide and Online Hostility

Research on digital participation has long documented significant disparities across online platforms. Contrary to early expectations that digital spaces would foster widespread civic participation (Papacharissi 2004), the "90-9-1" principle suggests that 90 percent of users passively consume content, 9 percent contribute occasionally, and only 1 percent generate the majority of online content (Nielsen 2006). Although comprehensive research on this inequality remains scarce, several studies confirm that only a small fraction of users actively participate in digital spaces (e.g., Van Mierlo 2014; Gasparini et al. 2020; Carron-Arthur, Cunningham, and Griffiths 2014; Baqir et al. 2023; Antelmi, Malandrino, and Scarano 2019).

The inequality of digital participation nevertheless remains largely unexplored. Most studies on the digital divide have focused on disparities in physical access to digital systems (Chaqfeh et al. 2023) or differences in digital skills and literacy (Hargittai 2018; Hargittai and Shaw 2015), with far less attention given to other dimensions of digital inequality (Korovkin, Park, and Kaganer 2023; Scheerder, Van Deursen, and Van Dijk 2017; Van Dijk 2006). Thus, there is limited understanding of the extent of participation inequality among individuals who have access to digital platforms but engage with them to varying degrees. This participation gap is especially salient in South Korea's portal-centered news environment, where access runs through centralized hubs that aggregate many outlets.

Prior research also suggests that digital participation inequality may be linked to a higher likelihood of hostile engagement. Hostility or incivility in online spaces has been widely documented, particularly in political discussions and news comment sections (e.g., Coe, Kenski, and Rains 2014; Humprecht, Hellmueller, and Lischka 2020; Rowe 2015; Santana 2014; Rossini 2022). In online comment sections, frequent users are more likely to post hostile content. For example, research on Facebook found that highly engaged users exhibit greater levels of toxicity in their comments (Kim et al. 2021a). Similarly, studies on news comment sections indicate that hostility tends to cluster among the most active participants (Humprecht, Hellmueller, and Lischka 2020; Rowe 2015), potentially shaping broader public perceptions of digital discourse.

The potential association between frequent commenting and hostile content may be driven by anger, a high-arousal emotion that is strongly linked to greater engagement and participation (Berger 2011; Brady et al. 2017; Crockett 2017; Hasell and Weeks 2016; Masullo, Lu, and Fadnis 2021; Valentino et al. 2011). This pattern is particularly pronounced in partisan digital environments, where hostility toward out-groups generates higher engagement than in-group favoritism (Rathje, Van Bavel, and Van Der Linden 2021; Yu, Wojcieszak, and Casas 2024). Masullo, Lu, and Fadnis (2021) further suggests that anger increases the likelihood of users actively expressing their opinions online, regardless of the opinion climate they encounter. In settings where comments are ranked by popularity or recency, these emotion–engagement dynamics can concentrate which voices and tones become most visible

Building on these insights, this study advances research on the digital divide by bridging two critical aspects of online engagement—digital participation inequality and online hostility—that have not been systematically examined together. By leveraging individual-level news comment behavior data over a 13-year period, this study provides a rare opportunity to examine both the severity of the participation divide between frequent and infrequent users and whether this divide is indeed linked to hostile engagement. Beyond showing that active users are more hostile, we examine how participation inequality relates to hostility across time and news domains by decomposing changes in overall hostility into composition effects (who is seen; shifts in group shares) and within-group rate effects (how hostile each group is) under the South Korean specific setting where centralized hubs aggregate many outlets and comments are ordered by popularity or recency.

Data

Naver News

South Korea is one of the most digitally connected countries in the world, boasting the highest percentage of high-speed broadband connections among OECD nations (Pak, André, and Beom 2021). In addition, in this country, online news consumption is overwhelmingly concentrated on news aggregator platforms rather than individual news websites. According to a global comparison of 46 countries, South Korea had the highest rate of news consumption via news aggregators and the lowest direct access to news websites in 2021 (Oh, Park, and Choi 2021). *Naver News* is the most dominant gateway in Korea—over 90 percent of Koreans use *Naver* as their primary search engine, and 87 percent rely on *Naver News* for their online news consumption (Kim et al. 2021b). This suggests that the inequality of digital access is at least minimal.

This context allows us to examine participation inequality without appealing to access constraints. *Naver's* in-link system lets users read full articles and comment within the platform, removing the need to create multiple outlet accounts and keeping news consumption and discussion in one place. During our study period, exposure on the news front pages was not shaped by personalized recommendation algorithms, reducing algorithmic bias in what users see. This comprehensive and centralized design makes *Naver* News particularly valuable for studying digital participation and hostile engagement at the individual level, with consistent definitions of exposure and discussion across outlets. With

longitudinal tracking of commenting at the user level, observed participation disparities are likely to reflect user preferences rather than structural access limitations.

News Comment Data

From Naver News, we collected approximately 260 million comments along with unique user identifiers (different from actual accounts; partially masked by the platform) from January 2008 to September 2020, using the R package *N2H4*. During this period, Naver News published a daily list of the 30 most-read articles across six domains—Politics, Society, Economy, World, IT/Science, and Life/Culture—amounting to 180 articles per day.

One important feature of the comment system is that it technically supports one level of nested replies, but is effectively semi-flat: most comments appear in a single top-level feed, and reply chains rarely extend beyond a few posts. Due to the limitations of the N2H4 collection tool, our dataset contains only the most visible top-level comments from the daily most-read list, rather than the full set of nested replies. While this was not an intentional design choice, it has the effect of aligning our analysis with the lurker experience, since casual readers are most likely to encounter these prominently displayed comments.

The final dataset contains 802,946 articles from 141 outlets and 260,203,552 comments posted by approximately 6,170,121 unique users. On average, each article received 324 comments. The volume of commenting activity increased significantly over time (see *Appendix*), likely reflecting the growing accessibility of online news platforms.

Given the large volume of comments and the presence of heavily active users, we conducted a supplementary analysis to assess whether certain users exhibited non-authentic (bot-like) behavior, such as repetitive commenting. We find that while a small subset of users show highly duplicated (or similar) content, their overall prevalence is limited and unlikely to bias the main results (see *"Check Bot-like Accounts"* subsection in the *Appendix*).

Hate Speech Data

Previous research distinguishes between intolerant messages, which express harmful or discriminatory intent toward specific groups, and impolite posts, which contain rude or offensive language (Rossini 2022). Although terminology varies across the literature (e.g., Rossini 2022; Rowe 2015; Rega, Marchetti, and Stanziano 2023), these studies emphasize that not all uncivil messages are equally damaging to democratic discourse, underscoring the need to differentiate between the two types. This study adopts a typology of "hostility" that distinguishes these forms: we refer to generally rude or offensive content as *uncivil*, and content targeting specific social groups as *hateful*.

To detect hostility in news comments, we adopt the *Korean Unsmile Dataset*(Kim 2022) as the labeled corpus to operationalize our hostility construct. How these labels are used to train and compare classifiers is described in *Measuring Comment Hostility*. This dataset includes ten distinct labels (civil, uncivil, and hate speech targeting nine different social groups), allowing us to evaluate model performance across nuanced types of hostile languages. The ten categories specifically include *Civil* (devoid of hate speech), *Uncivil* (disparaging language or personal attacks), and various hate speech types targeting *race/nationality*, *region*, *gender*(male/female/family), *religion*, *age*, and *sexual minorities*. Each comment can receive multiple labels across categories.

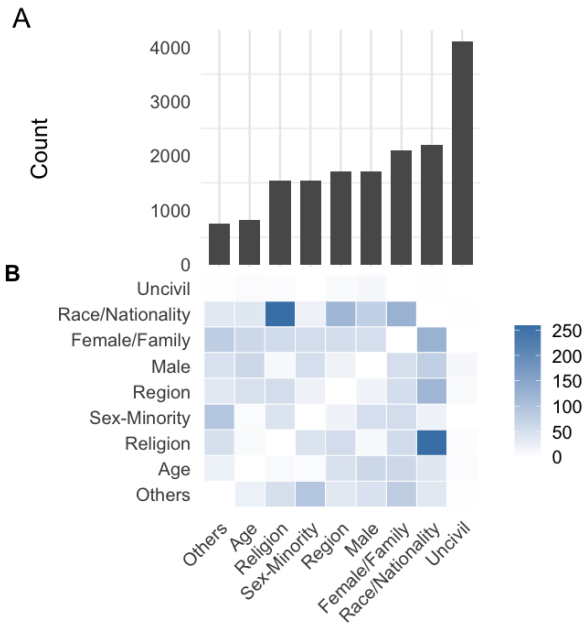


Figure 1: Label distribution and co-occurrence (multi-label).
(A) Marginal frequencies of Uncivil and hate categories; 'Civil' is excluded.
(B) Pairwise co-occurrence among labels (darker = greater overlap), showing how labels appear together within the same sentence.

However, this dataset may potentially misclassify neutral comments as hateful (Kang et al. 2022). For example, a benign statement referencing a group may be flagged as hate speech incorrectly. We supplemented the dataset with additional neutral sentences following Kang et al. (2022) to mitigate this issue. Despite these mitigations, Korean online discourse still employs obfuscation and morphology that can mask hostility—e.g., consonant-only abbreviations, derivational pejoratives, spacing/morphological variation, and code-mixing. Such devices degrade tokenization so that even a fine-tuned deep model can miss euphemistic/-coded hostility (false negatives) or flag reclaimed/sarcastic usage (false positives). We provide marked examples with the salient linguistic features in Table 8 in the *Appendix*.

In the training dataset, uncivil content is the most frequent category (24.5%), followed by hateful content targeting race/nationality (13%), female/family (12%), male (11%), region (10%), religion (9%), sex minority (9%), and age (4.8%). These frequencies are shown in Figure 1A. While panel Figure 1A shows label-wise frequencies, it does

not reflect how many comments carry multiple labels. Figure 1B visualizes pairwise co-occurrence among categories (darker cells indicate greater overlap), allowing readers to see at a glance both the marginal distribution (Figure 1A) and the overlap structure (Figure 1B) in a single integrated figure.

Methods

Measuring Participation Inequality

To assess user engagement levels, we first ranked all users in the dataset based on the number of comments they posted, with the most active commenters placed at the top. This ranking allowed us to classify users into different engagement groups, which were then used to compare hostility levels in their comments. Our analysis primarily focuses on the top 10% of the most active commenters, comparing them to the bottom 40% of commenters, who consistently exhibit substantially lower engagement.

To quantify participation inequality among these user groups, we employed two widely used economic disparity metrics: the Gini index and the Palma index (Atkinson et al. 1970; Kakwani 1977), both of which have been applied in prior research to assess engagement inequalities in digital spaces (Glenski, Volkova, and Kumar 2020).

The Gini index captures the overall dispersion of participation levels, reflecting how unequally comments are distributed among users. A higher Gini index indicates greater inequality in engagement. However, it has limitations: two distributions can share the same Gini value yet differ in whether the disparities are driven by the most or least active users. Moreover, it is more sensitive to changes in the middle of the distribution than at the extremes.

To address these limitations, the Palma index focuses on extremes by measuring the ratio of participation between the top 10% and the bottom 40% of commenters. An increasing Palma index indicates growing dominance of the most active users over the least active ones, offering a clearer picture of who dominates the discourse in digital spaces and to what extent.

Therefore, throughout our analysis, we present Gini and Palma side by side: Gini for overall inequality, and Palma for concentrating among highly active users. This complementary approach allows us to identify whether observed trends are driven by shifts in mid-level participation or by changes at the extremes.

Measuring Contribution to Inequality

After calculating the inequality metrics, we assess whether the observed disparities are primarily driven by frequent or infrequent commenters using the relative mean deviation (RMD). This metric is mathematically defined as follows:

$$RMD_{ig} = \frac{N_i - \mu_g}{\mu_g} \quad (1)$$

where i represents an individual user, g denotes the news domain. N_i is the number of comments posted by user i, and μ_g represents the average number of comments per user in news domain g.

The RMD serves as a counterfactual measure to evaluate participation inequality. In a scenario where all users contributed an equal number of comments, the comment space would exhibit perfectly equal participation. This hypothetical equal participation level is represented by μ_g. By comparing each user's actual comment count to μ_g, the RMD quantifies how much more or less each user contributes relative to this counterfactual equality.

This metric allows us to determine whether inequality is driven by frequent commenters posting significantly more than expected or by infrequent commenters contributing far less than the counterfactual amount. In doing so, it provides a clearer picture of how participation disparities emerge in online discussions.

Measuring Comment Hostility

Basic Framework To assess the level of hostility in user comments, we conducted a content analysis by stratifying commenters into heavy (top 10%) and light (bottom 40%) engagement groups, based on the Palma index. Within the top 10%, we further isolated the top 1% of commenters, as a small subset appeared disproportionately frequently compared to others.

As an initial step, we used *KC-BERT* (Base and Large) and *KC-Electra* (Lee 2020) models using the hate speech dataset described earlier. These KC-specific models are designed to better capture the nuances of Korean online comments, including informal variations and synonymous expressions. To benchmark their performance and assess classification robustness, we also trained two widely used reference models: *KoBERT* and *KoElectra*.

After model training, we selected the best-performing model and applied it to a 1% stratified sample of comments from each user group. The model assigned multi-label scores to each comment, and for simplicity, we retained only the highest-scoring label per comment, filtering out those for which all scores were below 0.5. We then consolidated the ten original hate categories into three broader classes: *civil*, *uncivil*, and *hateful*. This grouping allows for a more precise comparison of hostility level across commenter groups by reducing label complexity while preserving key semantic distinctions.

Specifically, *uncivil* comments include general profanity and personal attacks, whereas *hateful* comments contain derogatory or discriminatory expressions targeted at specific social groups (e.g., gender, religion, race, or region). Comments lacking such content are labeled *civil*. Based on this three-way classification, we then compared the distribution of component types (*civil*, *uncivil*, and *hateful*) across user engagement groups. Using a chi-squared test of proportions, we tested whether the observed differences in hostility levels between heavy and light commenters were statistically significant.

Details of the Training Process To determine optimal performance, we conducted experiments across a range of hyperparameters: learning rates of 2e-5, 3e-5, and 5e-5; batch sizes of 8, 16, 32, and 64; and training epochs of 3 and 5. All models were trained and evaluated using a sin-

gle NVIDIA Tesla T4 GPU on Google Colab. The training for each model took approximately 2–4 hours, depending on model size and batch configuration.

Table 1 summarizes the best-performing configurations. Among all models, *KC-Electra* slightly outperformed the others. This result aligns with our expectations, as KC-specific models are pretrained on Korean corpora that include substantial amounts of informal, user-generated content, such as online comments, making them more attuned to the linguistic characteristics of our dataset.

Model performance was evaluated primarily using the Label Ranking Average Precision (LRAP) score. Additional metrics—including precision, recall, and F1 Score—also confirmed the superior performance of KC-based models in handling the multi-label classification of hostile and hate speech in Korean text.

Model	F1	Precision	Recall	LRAP
KC-BERT Base	0.862	0.876	0.850	0.926
KC-BERT Large	0.865	**0.884**	0.851	0.928
KC-Electra	**0.872**	0.882	**0.866**	**0.931**
KoBERT	0.839	0.837	0.843	0.908
KoElectra	0.849	0.847	0.852	0.913

Table 1: Performance comparison of transformer-based models on the multi-label hate speech classification task. The best-performing model for each metric is shown in bold.

Participation Inequality

Descriptive statistics on participation levels indicate a stark digital participation gap (Figure 2). On average, the top 10% of frequent commenters account for nearly half of all comments in news comment sections (50.11%), while the bottom 40% contribute only 14.99% of total comments over the years. Importantly, this divide is not merely transient: cohort-based continuous retention shows that 64.2% of Top-10 users in a given month remain Top-10 in the next month, and 13.1% remain Top-10 for the entire next year. Together, Figure 2 and these persistence diagnostics indicate a stable concentration of participation and substantial divide in digital participation (full retention table and the survival plot in *Appendix*, Figure 7, Table 12). This imbalance underscores the motivation for our study, highlighting the need to investigate the structural disparities in online engagement.

Participation Inequality by News Domain and Popularity

To further examine this divide, we quantified participation inequality within the news ecosystem using the Gini index and the Palma index. We then compared participation inequality (a) across six news domains (*Politics, Society, Economy, World, IT/Science, and Life/Culture*) and (b) at varying levels of news popularity. Note that *Naver News* publishes a daily list of the 30 most-read articles, referred to as '*Ranking News*.' To measure news popularity, we used these rankings, with 1st representing the least popular and 30th the most popular article of the day. We then calculated Gini and Palma indices for different news stories based on their popularity ranks to assess how inequality changes across news interest levels.

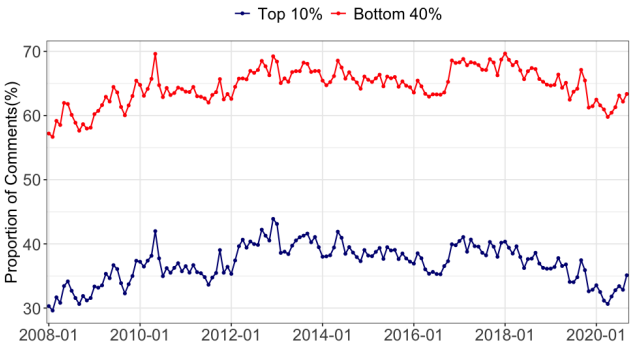


Figure 2: Share of Comments by Top 10% and Bottom 40% Groups

Figure 3 presents participation inequality across different news domains over the study period. From the Gini index, Politics emerges as the domain with the highest overall dispersion of participation, followed by Society and Economy, indicating that comments are highly concentrated among a limited number of users. The Palma index reinforces this pattern of Politics, showing its extreme concentration is driven primarily by the top 10% of commenters dominating over the bottom 40%. In contrast, domains such as Life/-Culture and IT/Science display relatively lower levels of the Gini index and markedly smaller Palma index, suggesting that lower overall inequality in these domains coincides with a more balanced participation between the most and least active users.

Taken together, the two measures indicate that while politics shows the highest overall inequality, its Palma values reveal that this pattern is largely driven by dominance among the most active commenters.

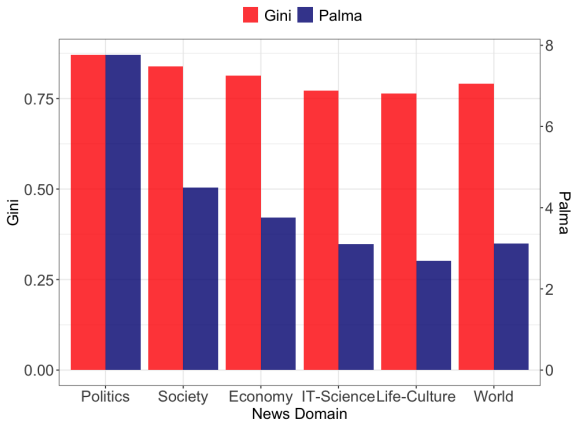


Figure 3: Average Gini and Palma Indices Across News Domains(2008-2020)

Figure 4 presents participation inequality as measured by the Palma index (Panel A) and the Gini index (Panel B) across different levels of news popularity. Across all domains, both indices display a clear upward trend, indicating that as a news story becomes more popular, participation inequality increases. The Palma index highlights that this effect is driven largely by the growing dominance of the most active commenters, while the Gini index confirms that overall disparities also widen with popularity. Domain patterns mirror the main trend: Politics/Society are highest and steepest on both Palma and Gini, whereas other domains change modestly.

At first sight, the finding that participation inequality becomes more pronounced in widely read articles may appear counterintuitive, since a larger readership could be expected to diversify participation. However, the semi-flat structure of the Naver News means that highly active users repeatedly occupy visible positions, especially in popular articles, thereby amplifying their dominance rather than diluting it. Combined with the absence of personalized ranking, this suggests that the heightened inequality observed in popular news reflects user behavior and platform design rather than algorithmic bias.

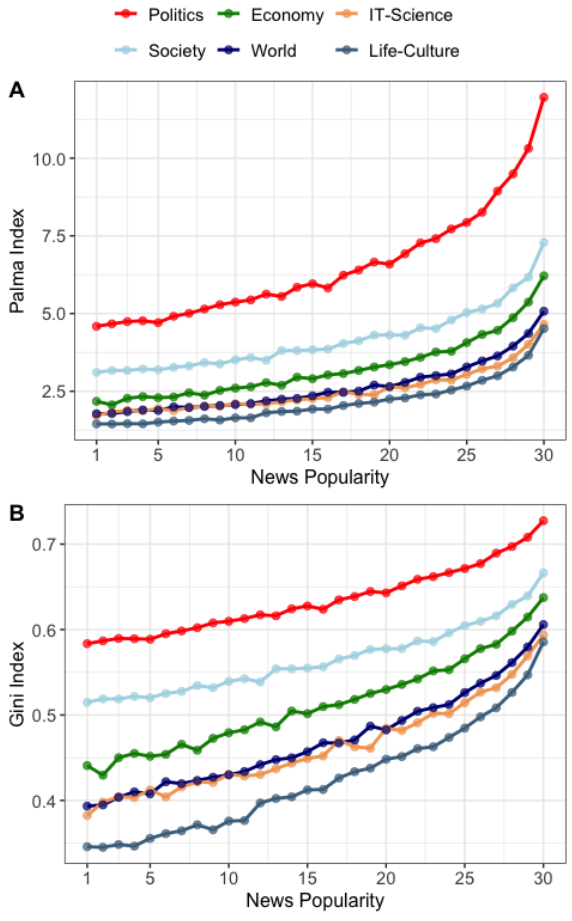


Figure 4: The Palma(Panel A) and Gini(Panel B) index by News Popularity

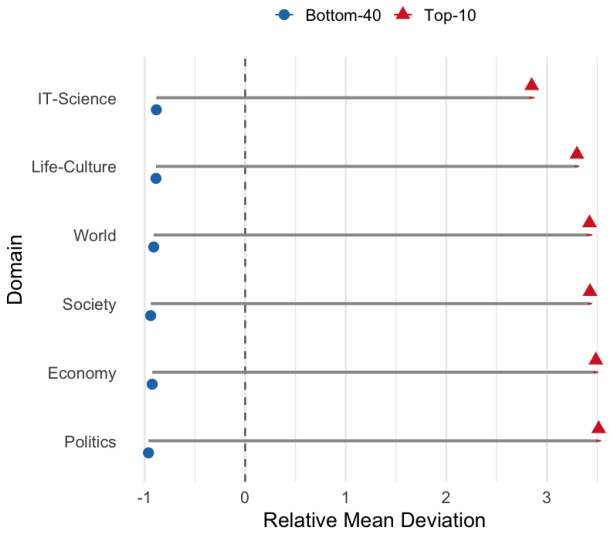


Figure 5: Relative Mean Deviation (RMD) by domain for Bottom-40 and Top-10 commenter groups by News Domain. RMD measures deviation from a counterfactual of equal participation. Negative values indicate under-participation; positive values indicate over-participation. For a finer breakdown within the top decile, see Figure 11 in the *Appendix*, which shows a sharp jump at the Top-1%.

User Contribution to Participation Inequality

To assess which user groups contribute most to participation inequality, we analyzed Relative Mean Deviation (RMD) scores. While the Palma and Gini indices measure overall inequality, they do not reveal how different user groups contribute to these disparities. RMD addresses this gap by indicating how much each group's participation deviates from a hypothetical benchmark of perfect equality, where all users contribute an equal number of comments within a given news domain and news popularity level. A value of 0 represents perfect equality, while negative values indicate lower-than-expected participation, and positive values indicate excessive participation relative to the equality benchmark.

Figure 5 presents the *relative mean deviation*(RMD)—the gap from a counterfactual of equal participation—for the bottom 40% and the top 10% of commenters across news domains. Across all domains, the bottom-40 sit slightly below zero (mild under-participation), whereas the top-10 lie well above zero (strong over-participation). The separation is largest in Politics and Society and smallest in IT-Science and Life-Culture.

Taken together, these patterns indicate that participation inequality is driven primarily by the over-contribution of the most active decline, rather than by a dramatic withdrawal of the least active 40%. This aligns with the Palma results (top vs. bottom concentration) and complements the Gini evidence on overall dispersion.

Moreover, there is a progressive and disproportionate increase in deviation among more active users, with the top 1% of commenters exhibiting the highest deviation. The top

1% of users have an RMD between 23 and 30, compared to an average deviation of 3 among other active groups, demonstrating their outsized influence on digital discourse (see Figure 11 in the *Appendix*).

These findings underscore two key aspects of participation inequality. First, they indicate that the observed participation gap is primarily driven by highly active users posting disproportionately more comments, rather than infrequent users posting significantly fewer comments. This suggests that participation inequality is a function of overcontribution by a small subset of users rather than disengagement by the majority.

Second, there is a sharp divide even among active commenters, particularly between the top 1% and the rest, highlighting that the most extreme contributors play a dominant role in shaping discussions. This suggests that online discourse is not only concentrated among a small subset of users but is further skewed by an even smaller group of hyperactive commenters, reinforcing the severe imbalances in digital participation.

Comment Hostility

Previous studies suggest that more active users in comment sections are more likely to exhibit hostility. To examine this, we conducted a computational content analysis to assess the levels of hostility in comments posted by different user groups. For this analysis, we focused on three distinct commenter groups, ranked by their commenting activity: (1) the top 1% most active commenters, (2) the next most active group (top 1–10% (excluding top 1%)), and (3) the bottom 40% least active commenters. It is important to note that the top 1% and top 1–10% are distinct groups, unlike the broader categories used in prior analyses.

Given the unique behavior of the most active users, as shown in the participation inequality results, we isolated the top 1% separately to better capture the extreme engagement patterns of this highly active subset. For each group, we randomly selected 1% of comments from the raw dataset for analysis. These comments were then classified as either (1) civil, (2) uncivil, or (3) one of eight types of hateful comments using a deep learning classifier trained on a large dataset of labeled comments.

Figure 6 presents the distribution of comment categories across these three user groups. For simplicity and clarity, we aggregated the original 10 fine-grained labels into 3 macro-categories—*civil*, *uncivil*, and *hateful*—and report these simplified proportions in Table 2, facilitating a concise comparison of hostility across groups.

As expected, the most frequent commenters—the top 1% and top 10%(excluding the top 1%)—are significantly more likely to post *uncivil* comments compared to the less active bottom 40%, as confirmed by chi-square proportion tests ($p < 0.001$ for the comparison between bottom 40% and top 1–10%, and between bottom 40% and top 1%).

To complement this finding, we computed Cohen's h (Cohen 1988), which indicated that these differences are not only statistically significant but also non-trivial in magnitude. All comparisons yielded $h \geq 0.2$, which meets the threshold for a small effect size (see *Appendix* for interpretation criteria), with the largest effect in the civil category ($h = 0.34$ for bottom 40% vs. top 1–10%; $h=0.42$ for bottom 40% vs. top 1 %), confirming that the observed differences—though driven primarily by shifts in civility rather than target-specific hate—are substantively meaningful rather than mere artifacts of large sample size. Together with the chi-square results, these findings provide strong evidence of behavioral divergence in commenting style across user activity levels.

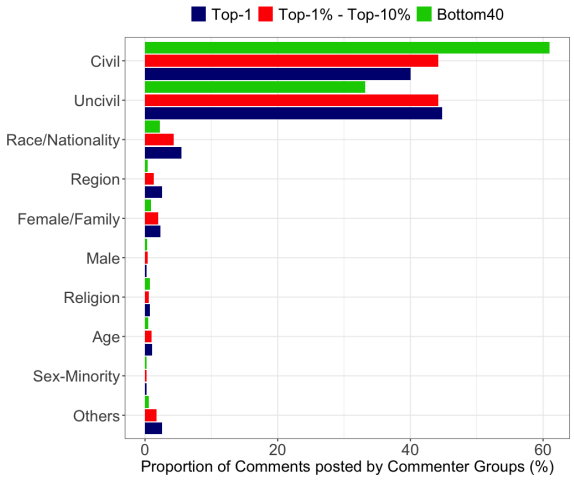


Figure 6: Hate Comment Classification Result by Percentile User Group

Regarding *hateful* content, the divide in online hostility extends even among active users: the top 1% is significantly more likely to post hateful comments than the top 1–10% (chi-square test, $p < 0.001$). This finding further reinforces the digital participation divide, showing that not only do a small number of users dominate discussions, but they also tend to engage in higher levels of incivility and hate speech.

User Group	Civil	Uncivil	Hate
Bottom 40%	0.610	0.332	0.058
Top 1% - Top 10%	0.442	0.442	0.116
Top 1%	0.400	0.448	0.152

Table 2: Proportion of comment categories by user group

Comparison	Civil	Uncivil	Hate
Bottom 40 vs Top 1% - Top 10%	0.34	0.23	0.21
Bottom 40 vs Top 1%	0.42	0.24	0.31

Table 3: Effect sizes (Cohen's h) across label categories by group comparison

We replicated the analysis using four alternative Korean language models to ensure that our classification results are

	Top 1%	Top 1–10% (excl. Top 1)	Bottom 40% (lower-base)
Civil	**65.1%**	33.0%	1.9%
Uncivil	**66.9%**	32.1%	1.0%
Hate	**73.4%**	26.0%	0.6%
Total n	1,336,325	635,515	28,919

Table 4: Composition of labels across activity groups (share %)

not overly dependent on model choice. The results consistently show that more active users tend to post a higher proportion of hate comments across all models. The full comparison of raw proportions and effect sizes (Cohen's h) is provided in *Robustness Check for Hate Speech Classification* in the *Appendix*

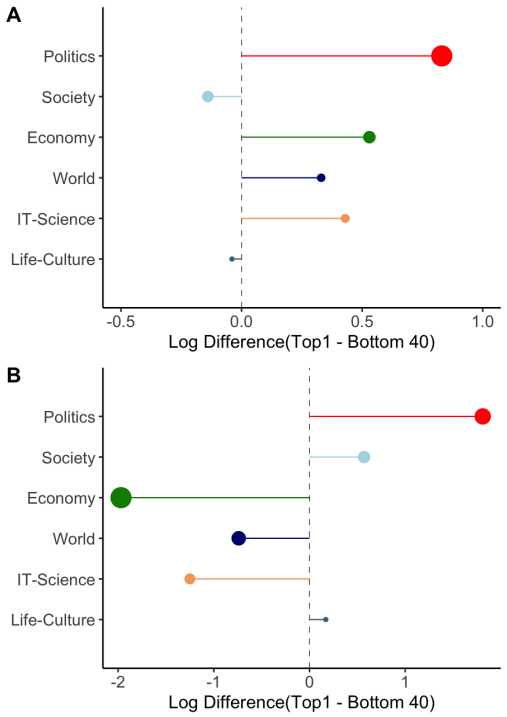


Figure 7: Log Difference in the Proportion of Uncivil (Panel A) and Hateful Comments (Panel B) between Extreme (Top 1%) and Inactive (Bottom 40%) User Group Across News Domains. Point sizes indicate the absolute difference in proportion.

The disparity in hostility between active and inactive groups is still evident when examining differences across news domains. As shown in Figure 7, the gaps in both uncivil and hateful comment proportions are particularly pronounced in the Politics domain, suggesting that highly engaged users are especially likely to contribute hostile discourse in political discussions.

Beyond the within-section analyses, we also provide an *across-group composition* view that asks which activity groups account for each label. Table 4 shows that output is dominated by upper-activity groups even within labels: the Top 1% alone accounts for **66.9%** of *Uncivil* and **73.4%** of *Hate*. Importantly, the Top 1% also dominates *Civil*—a consequence of their outsized *volume*—so composition shares alone can blur the tone-specific pattern.

Linking Participation and Inequality

Our descriptive results so far show that participation is highly concentrated. Our exploratory hostile-content classification points in the same direction: the most active users (the Top 10%, and especially the Top 1%) post hostile content at higher rates than the Bottom 40%, both across and within group summaries. The question now is whether the person who talks the most also shapes the tone that becomes visible in a systematic way.

To make that link concrete, we align the two measures on the same activity group. Inequality is captured by the Palma ratio—the share of the Top 10% relative to the Bottom 40%. Hostility is summarized with a group-conditional index that tracks three groups in each section-month: the Top 1%, the Top 1-10% (excluding Top 1%), and the Bottom 40% (used as a lower-base group). Looking at the same groups lets us focus on the comments most likely to shape what readers actually see when activity is concentrated, while providing a stable baseline from the large mass of infrequent commenters.

When a small set of users supplies most comments, the parts of the discussion that people actually see tend to reflect their style—often sharper and more confrontational—so simply increasing participation volume does not guarantee a broader or more civil conversation; indeed, a harsher tone can push casual users away, leaving the floor even more concentrated and reinforcing the same tone.

From this logic, we proceed in two steps. **(H1)** We relate concentration to hostile tone within sections over time. **(H2)** We then ask why hostility changes—separating shifts in who is talking how much (composition) from shifts in how the same groups talk (behavior)—using a simple change decomposition.

H1: Concentration and User Group-Conditional Hostility

Hypothesis. Sections with higher participation concentration (higher Palma) exhibit higher group-conditional hostility $H_{s,t}$

Measures. Group–conditional hostility, $H_{s,t}$ is defined as follows:

$$H_{s,t} = \sum_{g \in \text{Top1, Top1-10Top1, Bottom40}} w_{g,s,t} p_{g,s,t}$$

where $w_{g,s,t}$ is group shares renormalized to sum to one within these three user groups and $p_{g,s,t}$ is the group-specific hostility rate.

Model.

$$H_{s,t} = \alpha \log(\text{Palma}_{s,t}) + \gamma_s + \tau_t + \epsilon_{s,t}$$

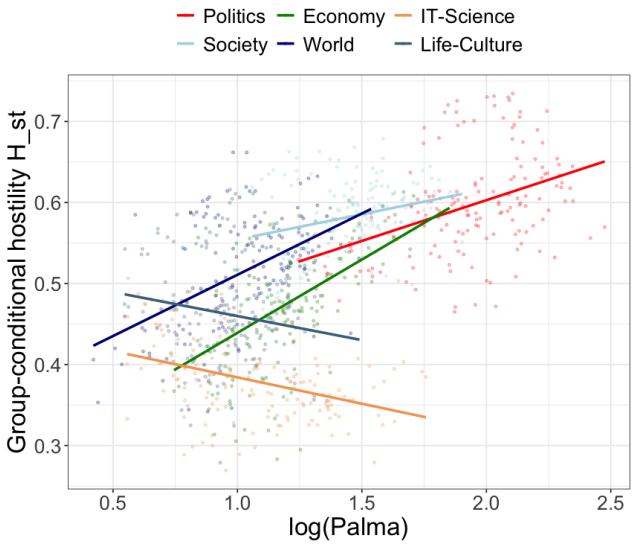


Figure 8: Relationship between group-conditional hostility $H_{s,t}$ and $\log(\text{Palma})_{s,t}$ by section. Lines show within-section linear fits.

The fixed-effects specification yields a positive and statistically significant association: $\hat{\alpha} = 0.056$(SE=0.016). With two-way clustered SE, the estimate is identical (0.056, SE = 0.015). Substantively, doubling the Palma ratio is associated with about 3.9 percentage points higher $H_{s,t}$. (Without section FE, the cross-sectional level differences inflate the coefficient to 0.145, SE = 0.031; our focus is on the within-section relationship.) Within a given section, months with a stronger concentration of participation correspond to higher hostility among the groups (Top 1%, Top1-10%, Bottom 40%). This pattern is consistent with the idea that "who speaks more" (concentration) moves the tone of discourse among the most consequential participants. The figure 8 and Table 5 report the details.

H2: Decomposing Changes in Hostility

Hypothesis. Month-to-month changes in group-conditional hostility within a section can be decomposed into **(i) composition** shifts among the groups (changes in group weights $w_{g,s,t}$ and **(ii) behavioral** shifts within groups (changes in hostility rates $p_{g,s,t}$). We quantify the relative contributions of these two channels.

Two-period$(t \to t+1)$ **Decomposition.** For each section s,

$$\Delta H_{s,t} \equiv H_{s,t+1} - H_{s,t} \approx \underbrace{\sum_g (\Delta w_{g,s,t}\, \bar{p}_{g,s,t})}_{\text{Composition}} + \underbrace{\sum_g \bar{w}_{g,s,t}\, (\Delta p_{g,s,t})}_{\text{Behavior}}.$$

where $\Delta x_{g,s,t} = x_{g,s,t+1} - x_{g,s,t}$ and $\bar{x}_{g,s,t} = \frac{1}{2}(x_{g,s,t} + x_{g,s,t})$. This yields a path-independent two-period decomposition and exactly sums to ΔH up to rounding.

	(1) OLS	(2) FE	(3) FE + 2-way SE
Dependent variable: $H_{s,t}$ (group-conditional hostility, 0–1)			
$\log(\text{Palma}_{s,t})$	0.145**	0.056*	0.056*
	(0.031)	(0.016)	(0.015)
Section FE	No	Yes	Yes
Month FE	No	Yes	Yes
SE clustered by	—	Section	Section & Month
Observations	918	918	918
R^2 (within)	0.63	0.91	0.91

Table 5: Concentration and group-conditional hostility (section–month panel).
Notes: Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. (3) reports two-way clustered SE.

Section	n pairs	Mean ΔH	Median Comp. Share	Median Behav. Share
IT-Science	140	0.0004	0.002	0.998
Economy	140	0.001	0.001	0.999
Society	140	0.001	0.005	0.995
Life-Culture	140	0.001	0.008	0.992
World	140	0.001	0.002	0.998
Politics	140	0.002	-0.001	1.001

Table 6: Decomposing month-to-month changes in group-conditional hostility by section.
Notes: Shares are medians over valid section–month pairs. Small rounding differences may occur.

Using this two-period decomposition, month-to-month changes in group-conditional hostility within each section are explained almost entirely by **behavioral channel** (changes in within-group hostility rates), while shifts in tail weights (composition) are negligible. Median behavioral shares are approximately 99-100% across sections(Table 6).

Conclusion

This study underscores the stark participation inequality in online news comment sections, where a small but highly active subset of users disproportionately shapes digital discourse. Analyzing 260 million comments over 13 years on *Naver News*, we find that this participation gap is particularly pronounced in political news discussions and highly popular news stories. The analysis also reveals that the most active commenters contribute disproportionately to the overall volume of engagement, further amplifying their influence. Moreover, these frequent commenters are significantly more likely to engage in hostile discourse, posting both uncivil and hateful content at higher rates than less active users. Building on this descriptive picture, we align inequality and hostility with the same activity groups and test two hypotheses. **H1** shows that higher participation concentration is *associated* with higher hostile tone within sections over time; **H2** indicates that changes in hostility are driven primarily by *behavioral* intensification among active users rather than shifts in who is present.

These findings carry important implications for digital

public discourse and online platform governance. When a few users supply most comments, the visible conversation can skew sharper and more confrontational, potentially discouraging casual participation and further concentrating voice. While our design is not perfectly causal and subject to confounding and measurement error, it *quantifies* a robust relationship previously noted only descriptively and clarifies a plausible visibility/attention mechanism linking concentration to hostile tone. Policy and design responses may therefore focus not only on broadening participation but also on shaping *how* dominant users engage.

Limitations

While this study provides valuable insights into digital participation inequality and hostile discourse, it has several limitations that should be addressed in future research.

First, although our findings reveal a significant disparity in hostility between active and inactive user groups, further analysis is needed to understand the underlying linguistic mechanisms driving this disparity. Specifically, a more granular examination of how hostile language is constructed and varies between these groups would provide deeper insights. However, this presents a methodological challenge due to the complex structure of the Korean language. Korean allows for the creation of new words through character combinations, often leading to non-standard lexical variations in online discussions. This makes tokenization particularly difficult, as conventional NLP methods may fail to capture these variations accurately.

Additionally, detecting hostility—especially hateful content targeting specific sociopolitical groups—is further complicated by implicit and coded expressions that may not contain overt hate speech terms but still convey derogatory or exclusionary meanings. This linguistic flexibility enables users to mask hostility, making deep-learning-based classification models prone to under-detection of such content. Addressing this issue requires more sophisticated linguistic processing techniques, such as context-aware tokenization models, morphological analysis tailored to Korean online discourse, and adversarial training methods that can better capture implicit hostility. Future research should refine these approaches to improve the precision of hostility detection, particularly for nuanced forms of incivility and hate speech.

Second, although we test hypotheses with regressions (with section and time controls), this does not identify causality; residual confounding may remain (e.g., contemporaneous topic shocks such as elections, changes in moderation/ranking UI, selective visibility by highly active users). Our dataset is limited to observational digital trace data, which primarily captures user behaviors, comment timing, and content, but does not account for underlying psychological or social motivations. Future research should explore experimental methods to better understand the causal links between participation inequality and online hostility.

Despite these limitations, this study provides a foundational analysis of how a small proportion of users shapes digital discourse through both disproportionate engagement and elevated hostility. The findings are particularly novel given the scale and granularity of the dataset, as well as the platform's minimal algorithmic bias—specifically, the absence of personalized ranking in the daily most-read articles—which allows for clearer attribution of behavioral patterns. This advances our understanding of how participation inequality can distort democratic discourse, even in relatively open digital environments.

Addressing these challenges in future research will be crucial for developing more effective moderation strategies and fostering healthier online discussions.

References

Antelmi, A.; Malandrino, D.; and Scarano, V. 2019. Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1035–1038. San Francisco USA: ACM. ISBN 978-1-4503-6675-5.

Atkinson, A. B.; et al. 1970. On the measurement of inequality. *Journal of economic theory*, 2(3): 244–263.

Baqir, A.; Chen, Y.; Diaz-Diaz, F.; Kiyak, S.; Louf, T.; Morini, V.; Pansanella, V.; Torricelli, M.; and Galeazzi, A. 2023. Beyond Active Engagement: The Significance of Lurkers in a Polarized Twitter Debate. ArXiv:2306.17538 [physics], arXiv:2306.17538.

Berger, J. 2011. Arousal increases social transmission of information. *Psychological science*, 22(7): 891–893.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Van Bavel, J. J. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Carron-Arthur, B.; Cunningham, J. A.; and Griffiths, K. M. 2014. Describing the distribution of engagement in an Internet support group by post frequency: A comparison of the 90-9-1 Principle and Zipf's Law. *Internet Interventions*, 1(4): 165–168. Publisher: Elsevier.

Chaqfeh, M.; Asim, R.; AlShebli, B.; Zaffar, M. F.; Rahwan, T.; and Zaki, Y. 2023. Towards a World Wide Web without digital inequality. *Proceedings of the National Academy of Sciences*, 120(3): e2212649120.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4): 658–679.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd edition.

Crockett, M. J. 2017. Moral outrage in the digital age. *Nature human behaviour*, 1(11): 769–771.

Gasparini, M.; Clarisó, R.; Brambilla, M.; and Cabot, J. 2020. Participation Inequality and the 90-9-1 Principle in Open Source. In *Proceedings of the 16th International Symposium on Open Collaboration*, 1–7. Virtual conference Spain: ACM. ISBN 978-1-4503-8779-8.

Glenski, M.; Volkova, S.; and Kumar, S. 2020. User Engagement with Digital Deception. In Shu, K.; Wang, S.; Lee, D.; and Liu, H., eds., *Disinformation, Misinformation, and Fake News in Social Media*, 39–61. Cham: Springer International

Publishing. ISBN 978-3-030-42698-9 978-3-030-42699-6. Series Title: Lecture Notes in Social Networks.

Hargittai, E. 2018. The digital reproduction of inequality. In *The inequality reader*, 660–670. Routledge.

Hargittai, E.; and Shaw, A. 2015. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, communication & society*, 18(4): 424–442.

Hasell, A.; and Weeks, B. E. 2016. Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Human Communication Research*, 42(4): 641–661.

Humprecht, E.; Hellmueller, L.; and Lischka, J. A. 2020. Hostile Emotions in News Comments: A Cross-National Analysis of Facebook Discussions. *Social Media + Society*, 6(1): 205630512091248.

Kakwani, N. C. 1977. Applications of Lorenz curves in economic analysis. *Econometrica: Journal of the Econometric Society*, 719–727.

Kang, T.; Kwon, E.; Lee, J.; Nam, Y.; Song, J.; and Suh, J. 2022. Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Aid Developing Better Hate Speech Dataset? arXiv:2204.03262.

Kim, J. W.; Guess, A.; Nyhan, B.; and Reifler, J. 2021a. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6): 922–946.

Kim, S. 2022. Korean UnSmile dataset: Human-annotated Multi-label Korean Hate Speech Dataset. https://github.com/smilegate-ai/korean_unsmile_dataset.

Kim, Y.; Shin, Y.; Sim, H.; Jang, Y.; and Mingyoo, P. 2021b. Media Users in Korea 2021. Technical report, Korea Press Foundation.

Korovkin, V.; Park, A.; and Kaganer, E. 2023. Towards conceptualization and quantification of the digital divide. *Information, Communication & Society*, 26(11): 2268–2303.

Lee, J. 2020. Kcbert: Korean comments bert. In *Annual Conference on Human and Language Technology*, 437–440. Human and Language Technology.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8): 707–710.

Masullo, G. M.; Lu, S.; and Fadnis, D. 2021. Does online incivility cancel out the spiral of silence? A moderated mediation model of willingness to speak out. *New Media & Society*, 23(11): 3391–3414.

Nielsen, J. 2006. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities. https://www.nngroup.com/articles/participation-inequality/. Accessed: 2024-01-06.

Oh, S.-U.; Park, A.; and Choi, J. 2021. Digital News Report in Korea 2021. https://www.kpf.or.kr/front/research/selfDetail.do?seq=592216.

Pak, M.; André, C.; and Beom, J. 2021. DIGITALIZATION IN KOREA: A PATH TO BETTER SHARED PROSPERITY? Technical report, Korea Economic Institute of America.

Papacharissi, Z. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2): 259–283.

Rathje, S.; Van Bavel, J. J.; and Van Der Linden, S. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26): e2024292118.

Rega, R.; Marchetti, R.; and Stanziano, A. 2023. Incivility in online discussion: An examination of impolite and intolerant comments. *Social Media+ Society*, 9(2): 20563051231180638.

Rossini, P. 2022. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, 49(3): 399–425.

Rowe, I. 2015. Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society*, 18(2): 121–138.

Santana, A. D. 2014. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice*, 8(1): 18–33.

Scheerder, A.; Van Deursen, A.; and Van Dijk, J. 2017. Determinants of Internet skills, uses and outcomes. A systematic review of the second-and third-level digital divide. *Telematics and informatics*, 34(8): 1607–1624.

Valentino, N. A.; Brader, T.; Groenendyk, E. W.; Gregorowicz, K.; and Hutchings, V. L. 2011. Election night's alright for fighting: The role of emotions in political participation. *The journal of politics*, 73(1): 156–170.

Van Dijk, J. A. 2006. Digital divide research, achievements and shortcomings. *Poetics*, 34(4-5): 221–235. Publisher: Elsevier.

Van Mierlo, T. 2014. The 1% rule in four digital health social networks: an observational study. *Journal of medical Internet research*, 16(2): e2966. Publisher: JMIR Publications Inc., Toronto, Canada.

Yu, X.; Wojcieszak, M.; and Casas, A. 2024. Partisanship on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2): 799–824.

Paper Checklist

1. For most authors..
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. In

the comment data, we explicitly address potential artifacts arising from population-specific distributions. For example, we examine the behavior of highly active users (top 1%) for possible bot-like patterns. We also perform stratified sampling to ensure that group proportions. See the Data and Appendix.

(e) Did you describe the limitations of your work? Yes. See the "Limitations"

(f) Did you discuss any potential negative societal impacts of your work? No, because our work focuses on characterizing the inequality and hostility observed in online news comment spaces. It does not involve any individual targeting or intervention that could directly impact users or platforms. We believe that the findings contribute to a better understanding of structural issues in digital participation and do not pose any foreseeable societal harms.

(g) Did you discuss any potential misuse of your work? Yes. In the "Limitations" section, we explicitly address the potential misinterpretation of our findings as causal claims.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. The user identifiers in the comment data are partially masked and pseudonymized by the platform. While these identifiers do not allow for re-identification of individual users, they remain unique and consistent across comments, allowing for longitudinal analysis of user behavior without compromising privacy. We do not have access to any raw account information, and no attempts were made to infer user identity.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA, While our study does not present theoretical results, we do discuss the implications of our empirical findings for understanding digital participation inequality and hostility.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the "Methods" part. The full dataset contains 260M raw comments, including potential PII, and is subject to the platform's Terms of Service. However, we provide a de-identified sample dataset and aggregated statistics required to reproduce the main regression results in the provided GitHub repository.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No. We did not repeat training with different random seeds, so error bars are not reported. However, we trained each model across multiple configurations of learning rates, batch sizes, and epochs, and selected the best-performing model based on validation metrics. This allows us to evaluate robustness with respect to hyperparameter sensitivity, although randomness in initialization was not explicitly tested.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, because our model is used only for group-level analysis rather than individual classification or moderation, we did not explicitly discuss the cost of misclassification. However, we acknowledge that false positives or negatives may affect estimated proportions, and we mitigate this risk through robustness checks across multiple models.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, we cited the creator of "Hate speech dataset"

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No. We used publicly available user comments from a major news platform. Although explicit consent was not obtained from users, the data is public, user identifiers are masked and anonymized, and no individual users are targeted or identified in the analysis. We use

the data solely for academic purposes and report results only at the aggregate level.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. The comment data contains no personally identifiable information; user identifiers are partially masked and pseudonymized by the platform, making re-identification impossible.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset ? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Descriptive Statistics for the Comment Dataset

Change in the Size of Comment Space The size of the comment space has grown rapidly over the years (Figure 9), and since our analysis focuses only on articles that received comments, we exclude users who did not engage in commenting. According to the widely cited 90-9-1 rule of online participation, approximately 90% of users typically consume content without contributing. This suggests that our observed comment-based measures of participation inequality likely provide a lower-bound estimate.

Distribution of the Number of Comments Online comment space is highly skewed. The Figure 10 shows a log-log(log-binned) distribution of comments per user, revealing a heavy upper tail: most users post only one or two comments, whereas a small minority produce orders of magnitude more. Given this scale disparity, percentile-based comparisons are more informative; accordingly, we standardize on the top 10% and bottom 40% groups throughout the paper.

Additional Details on Participation Inequality

Measurement and Grouping Each month, t, we rank users by the number of comments they post that month (descending). Ties at percentile boundaries are broken deterministically by the user's ID, ensuring reproducibility. We then form mutually exclusive activity groups on a rolling monthly basis: Top 1%, Top 1-10% (excluding Top 1%), and Bottom 40%. Unless noted, figures in the main text and below report Top 10% (Top 1% + Top 1-10%) and Bottom 40% as the focal comparison, in line with Palma-style summaries. For persistence analyses, groups are defined at the user-level across sections (i.e., users' total monthly comments), because the question is whether the same users remain highly active over time.

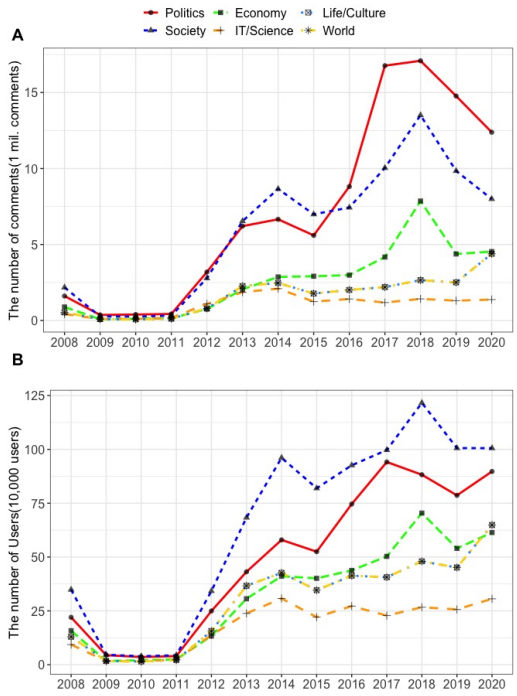


Figure 9: Change in the size of comment space: A. Change in the number of comments over time. B. Change in the number of users over time

Expanded Percentile Breakdown To diagnose where participation inequality concentrates within the activity distribution, we expand the RMD analysis from the main text's Top-10 vs Bottom-40 comparison to ten percentile bins (Bottom-10% ... Top-1%). Figure 11 shows a monotonic rise with a sharp jump at the very top: the Top-1% accounts for a disproportionately large deviation even relative to the rest of the Top-10%. This confirms that the main-text Top-10 effect is concentrated at the extreme tail rather than being uniform within the upper decile.

Persistence of Activity Ranks (Retention and Survival) To test whether participation concentration is transient or structurally sticky, we quantify persistence with two complementary diagnostics that use monthly counts only.

Continuous Retention. For each month t, let C_t be the set of users in the Top 10%. A user from C_t is considered retained at horizon k if they remain in the Top-10 in every month from $t+1$ through $t+k$. We summarize one-, three-, six-, and twelve-month horizons (reported as cohort-weighted averages across all eligible months). These statistics answer: "If a user is Top-10 in month t, what is the

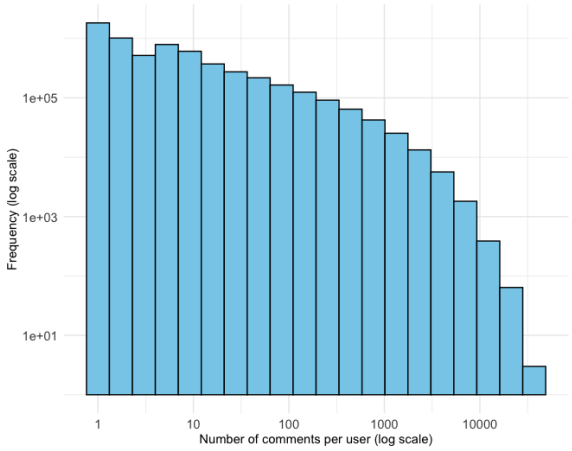


Figure 10: Distribution of comments per user on log-log axes (log-binned).

probability they are still Top-10 continuously over the next k months?"

Time to First Exit. We also estimate a Kaplan-Meier survival curve for the length of the first continuous Top-10 tenure per user (measured from the first observed Top-10 month until the first month they drop below Top-10). Tenures that reach the sample end are treated as right-censored; re-entries start a new tenure and are not used in the first-exit analysis. The survival curve reports the median Top-10 tenure and the probability of remaining in the Top-10 beyond specific horizons (e.g., 6 or 12 months).

Conclusion Together, the retention table and the survival curve show high short-run persistence and a non-trivial long tail (a durable core of highly active users), clarifying that the participation gap documented in the main text is not driven solely by short-lived fluctuations.

Horizon	Kept	Cohort	Retention
1 month	2,272,875	3,541,022	0.642
3 months	1,325,504	3,433,902	0.386
6 months	795,330	3,298,487	0.241
12 months	397,313	3,026,344	0.131

Table 7: Continuous Top-10 retention (cohort-weighted). One-, three-, six-, and twelve-month continuous retention rates with cohort totals (kept/cohort).
Note. A user in the Top-10 at month t counts as "kept" at horizon k only if they remain Top-10 in every month up to $t+k$. Values are averaged across all eligible cohorts (months with $t+k$ observed).

Check Bot-like Accounts

Frequent commenters exhibit a range of behavioral patterns. While some highly active users contribute large volumes of comments across multiple news domains, others post large quantities of near-identical content repeatedly. To assess the extent of such anomalous behavior, we analyzed duplication and similarity scores among users in the frequent commenter group.

For this analysis, we used the same stratified sample originally constructed for hostility classification. We focused on extremely active users (top 1%), as bot-like accounts are more likely to be concentrated in this group. In contrast, users in the bottom 40% group typically posted only once or twice, making the presence of bots in that group highly unlikely.

To measure the degree of similarity between user comments, we use the Levenshtein distance (Levenshtein 1966). This method calculates the minimum number of operations—such as deletion, insertion, and substitution—required to transform one string (*string A*) into another (*string B*). It provides a simple yet effective way to quantify textual similarity between pairs of comments. The distance is mathematically defined as follows:

$$\text{lev}(i,j) = \begin{cases} \max(i,j), & \text{if } min(i,j) = 0 \\ \min \begin{cases} \text{lev}(i-1,j)+1 \\ \text{lev}(i,j-1)+1 \\ \text{lev}(i-1,j-1)+\delta(a_i,b_j) \end{cases} & \text{o.w} \end{cases} \tag{2}$$

where

$$\delta(a_i,b_j) = \begin{cases} 0 & \text{if } a_i = b_j, \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

Based on the number of duplicated comments and Levenshtein distance, we defined the rate of duplicated comments and the similarity score of user i as follows:

$$Dup_i = 1 - \frac{No.\ Duplicated\ Comments_i}{No.\ Total\ Comments_i} \tag{4}$$

$$Sim_i = 1 - \frac{\sum_{a_i,b_i \in c_i} \text{lev}(a_i,b_i)}{No.\ Total\ Comments_i} \tag{5}$$

where a_i and b_i are distinct comments that user i posted and C_i is the collection of all comments that user i posted.

We find that, although a small subset exhibits highly repetitive behavior, the vast majority of users do not engage in abnormal posting patterns such as frequent comment duplication(Figure 13). This suggests the presence of potentially bot-like accounts in the comment space. However, their prevalence appears to be limited and unlikely to significantly distort the overall patterns of participation or hostility observed in our analysis.

Further Details about Hate Speech Dataset

Overall Structure of Dataset While the main text presents the distribution of individual hate speech labels, the dataset is broadly divided into two overarching categories: **Civil** and **Uncivil + Hate**. These two groups occur with nearly equal frequency, highlighting that the dataset is balanced in terms of overall tone.

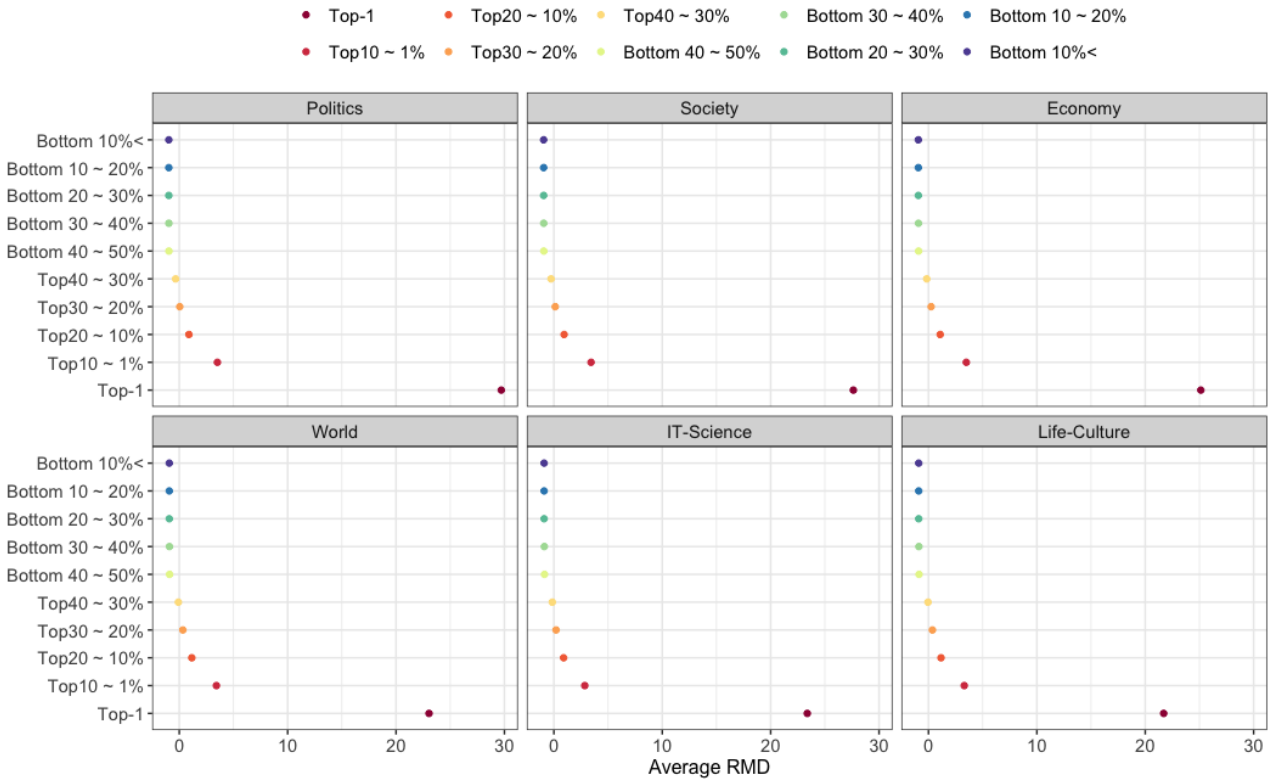


Figure 11: Relative Mean Deviation (RMD) by ten percentile bins(across news domains).

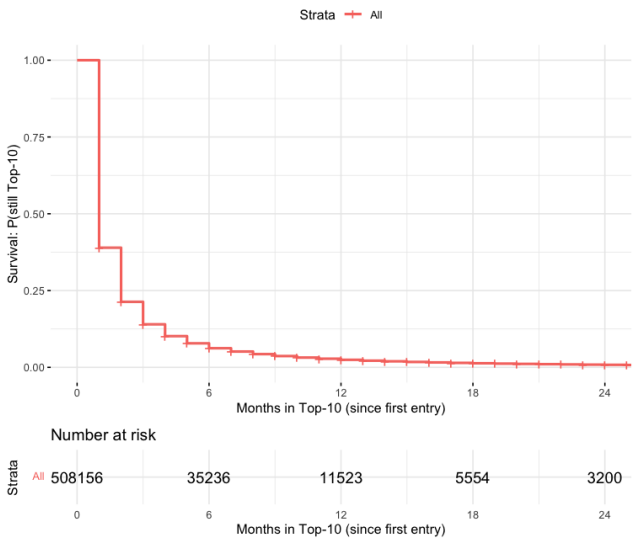


Figure 12: Time to first exit from the Top-10 (Kaplan–Meier). Tenures are measured from the first observed Top-10 month; tenures that reach the sample end are right-censored; re-entries start new tenures.

Overlap among Uncivil and Hate Categories Given the multi-label nature of the dataset, a single comment can exhibit multiple forms of hostility. To better understand such overlaps, we visualize a label co-occurrence heatmap in Figure 1B, where each cell represents the number of comments assigned to both corresponding labels.

Notably, *religion* and *race/nationality* often appear together, as do *region* and *race/nationality*. These patterns indicate the presence of overlapping hostile narratives targeting multiple social groups within individual comments.

One thing to note is that while each comment in our dataset may be assigned multiple labels, and some overlapping patterns are observed, the average label cardinality is relatively low at 1.1, suggesting that most comments are tagged with only a single category.

Example of Comments(Hateful)

Why detection is hard. Korean online text frequently uses (i) *consonant-only abbreviations*, (ii) *derivational pejoratives*, (iii) *spacing/morphological variation* that shifts token boundaries, and (iv) *code mixing/phonetic loans*. These devices convey derogatory or exclusionary meaning without overt slur tokens, degrading off-the-shelf tokenization and lexical cues.

Measurement implications. We rely on a fine-tuned *KC-Electra* multi-label classifier rather than surface-term rules, but the phenomena above still induce *false negatives* for eu-

Label	Korean Original (English Rendering)	Linguistic Feature
Female/Family	"한국ㄴ들 여학교 학교폭력 장난아님 여자들끼리 패거리만들고 왕따시키고 인간 아님." (*"Girls' high schools are full of cliques and bullying; they are not humane."*)	Abbrev./slur-like letter for '여자' ("ㄴ"), fragmented syntax
Male	"ㅇㄱㄹㅇ 의부심 부리는 것들 백프로 남자임" (*"Those who boast about being overly authoritative are all men."*)	Abbrev. ㅇㄱㄹㅇ (= "이거 레알" / "for real"), phonetic English loan, slang
Sexual Minority	"동성애는 사회에 혐오스러운 행위" (*"Homosexuality is viewed as a repulsive act in human society."*)	Hyperbolic rhetoric
Race/Nationality	"그중에서도 [MASK:SLUR]라니 하... 중국인은 참 미개하다" (*"[Slur]... Chinese people are uncivilized."*)	Ethnic slur (masked), onomatopoeic sigh "하...", informal particles
Age	"꼰대들이 없어져야 성인지 감수성이 개선된다" (*"Older generations must disappear for gender sensitivity to improve."*)	Culture-specific slang "꼰대" (generational insult)
Region	"경상도... 경기도... 강원도... 충청도!! 멍청도!" (*"People from certain regions are mocked as unintelligent."*)	Wordplay on province names; abusive rhyming/phonetic distortion
Religion	"종교[MASK:-충]들은 다 없어져야" (*"People of certain religions should not exist."*)	"-충" derivation (insect/parasite metaphor for groups)

Table 8: Masked examples of hostile language and salient linguistic features

User Group	KC-BERT Base			KC-BERT Large			KoBERT			KoElectra		
	Civil	Uncivil	Hate	Civil	Uncivil	Hate	Civil	Uncivil	Hate	Civil	Uncivil	Hate
Bottom 40%	0.674	0.258	0.068	0.705	0.226	0.069	0.669	0.247	0.084	0.631	0.305	0.064
Top 1% - Top 10%	0.524	0.341	0.135	0.550	0.311	0.138	0.546	0.316	0.138	0.493	0.386	0.121
Top 1%	0.486	0.337	0.177	0.506	0.303	0.191	0.511	0.318	0.171	0.455	0.389	0.156

(a) Raw proportions of comment categories by user group

Comparison	KC-BERT Base			KC-BERT Large			KoBERT			KoElectra		
	Civil	Uncivil	Hate	Civil	Uncivil	Hate	Civil	Uncivil	Hate	Civil	Uncivil	Hate
Bottom 40% vs Top 1% - Top 10%	0.310	0.182	0.225	0.322	0.192	0.230	0.253	0.153	0.173	0.279	0.081	0.057
Bottom 40% vs Top 1%	0.383	0.173	0.341	0.410	0.175	0.373	0.323	0.158	0.265	0.355	0.177	0.301

(b) Cohen's h for civil, uncivil, and hateful categories

Table 9: Comparison of raw comment proportions and Cohen's h across models and user groups

phemistic/coded hostility and *false positives* for reclaimed or sarcastic uses. Hostility rates in the main text should therefore be read as *model-based estimates* subject to under-/over-detection at the margin.

Illustrative examples. Table 8 presents masked snippets by target label (gender, age, region, etc.) alongside the salient linguistic features. We retain orthographic quirks in the Korean originals and provide concise English renderings for clarity.

Robustness Check for Hate Speech Classification

We conducted additional analyses using various pretrained language models to ensure our classification results are not model-specific. As shown in Table 9, the core pattern persists across all models: the most active users consistently exhibit a higher proportion of hate comments than less active users. Although the magnitude of group differences varies slightly, as indicated by Cohen's h, the direction and relative scale of the differences remain stable, supporting the robustness of our main findings

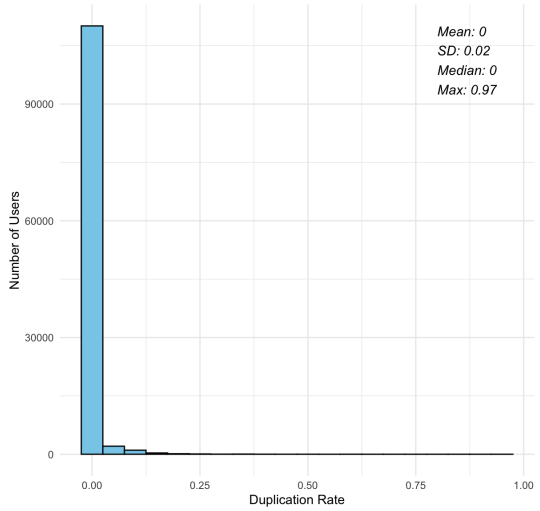
Cohen's h

Cohen's h (Cohen 1988) is a measure of effect size for differences between two proportions, defined as:

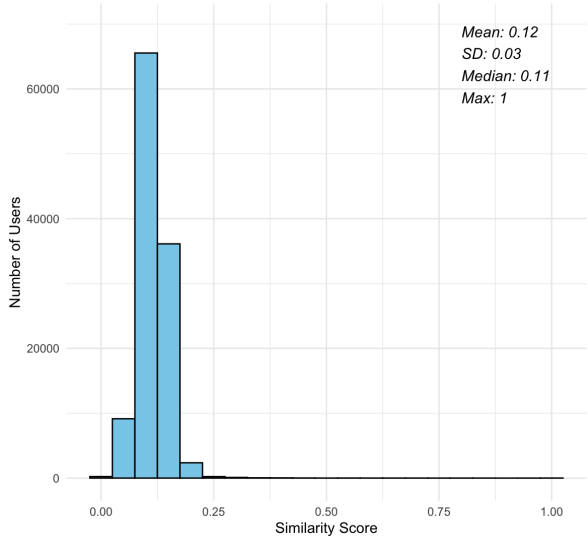
$$h = 2 \cdot \arcsin(\sqrt{p_1}) - 2 \cdot \arcsin(\sqrt{p_2})$$

The thresholds for interpreting h are as follows:

- $h = 0.20$: small effect size
- $h = 0.50$: medium effect size
- $h = 0.80$: large effect size



(a) Duplication rate distribution among frequent users.



(b) Similarity score distribution using Levenshtein distance.

Figure 13: Exploratory analysis of user behavior: (a) exact duplication rates and (b) content similarity.