The Loud Minority: How a Few Frequent Commenters Shape Digital Discourse with Hostility

Anonymous ACL submission

Abstract

Digital platforms were expected to foster broad participation in public discourse, yet online engagement remains highly unequal and underexplored. This study examines the digital participation divide and its link to hostile engagement in news comment sections. Analyzing 260 million comments from 6.2 million users over 13 years on *Naver News*, South Korea's largest news aggregation platform, we quantify participation inequality using the Gini and Palma indexes and estimate hostility levels with a BERT-based deep learning model. The findings reveal a highly skewed participation structure, with a small group of frequent users dominating discussions, particularly in Politics and Society and widely read stories. Participation inequality spikes during presidential elections, and frequent commenters are significantly more likely to post hostile content, suggesting that a vocal, and often hostile, minority disproportionately shapes digital discourse. By leveraging individual-level digital trace data, this study provides empirical insights into the behavioral dynamics of online participation inequality and its broader implications for digital public discourse.

1 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 et al., 2014; Baqir et al., 2023; Antelmi et al., 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 indexes to quantify participation disparities and apply a BERT-based deep learning model to classify comment hostility levels.

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 a more widely read news stories. Notably, participation inequality spikes during presidential elections, suggesting that major political events exacerbate engagement disparities. Moreover, these frequent commenters are significantly more likely to post hostile content, including both uncivil and hateful content, indicating that digital discourse is shaped by a vocal, and often hostile, minority.

By leveraging individual-level digital trace data, which offers a rare opportunity to observe engagement disparities at a granular level, this study provides empirical insights into the behavioral mechanisms underlying digital discourse inequalities and their broader implications for online public discourse and public opinion formation.

2 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 et al., 2014; Baqir et al., 2023; Antelmi et al., 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 et al., 2023; Scheerder et al., 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.

Prior research also suggests that digital participation inequality may be linked to a higher likelihood of hostile engagement. Hostility or incivility in online spaces have been widely documented, particularly in political discussions and news comment sections (e.g., Coe et al., 2014; Humprecht et al., 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 et al., 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 et al., 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 et al., 2021; Yu et al., 2024). Masullo et al. (2021) further suggests that anger increases the likelihood of users actively expressing their opinions online, regardless of the opinion climate they encounter.

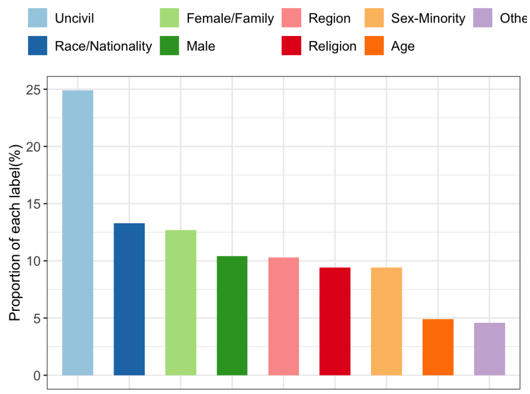

Figure 1: Distribution of Hateful and Uncivil Sentences. ’Civil’ sentences are excluded in this figure. We allow for overlapping counts here. If a sentence has two labels, it will be counted once for each label

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.

3 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 et al., 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 rate via direct access to news websites in 2021 (Oh et al., 2021). Among these platforms, *Naver News* stands as the most dominant, reflecting its unparalleled role in shaping the country’s digital news ecosystem. 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 shows the inequality of digital access is at least very little at play.

This minimal digital access inequality ensures that disparities in online engagement are not driven by differences in basic access to digital infrastructure but rather by individual preferences and behavioral choices. Unlike in countries where digital divides are primarily shaped by disparities in internet access, South Korea presents a unique context where virtually all users have the opportunity to engage with news content online, allowing for a more precise examination of participation inequality in digital discourse.

The platform, *Naver News* offers users free access to news content from major news outlets in the country. A key feature of the platform is its in-link system, which enables users to read full articles and comment on them directly within *Naver*, rather than being redirected to the original news websites. This design eliminates the need for users to create accounts on multiple media sites, effectively centralizing news consumption and discussion within a single platform.

The comprehensive scope of *Naver News* and its centralized commenting system make its data particularly valuable for studying digital participation and hostile engagement at the individual level. Because South Korea has minimal barriers to internet access, participation disparities on the platform likely reflect user preferences rather than structural access limitations. Moreover, *Naver News* data allows for tracking individual commenting behavior over time, providing a rare opportunity to examine participation patterns based on frequency of use.

News Comment Data

From Naver News, we collected approximately 260 million comments along with unique user identifiers from January 2008 to September 2020. During this period, *Naver News* published a daily list of the top 30 most-read articles ("Ranking News") across six news domains: Politics, Society, Economy, World, IT/Science, and Life/Culture, totaling 180 articles per day. The dataset comprises 802,946 articles from 141 news outlets, with 260,203,552 comments posted by approximately 6,170,121 unique users. On average, each article received 324 comments.

Hate Speech Data

To classify hostility in news comments, we trained a BERT based deep-learning model using the *Korean Unsmile Dataset*, a hate speech dataset provided by Smilegate-AI '(Kim, 2022). The dataset defines hateful expressions as those involving hostile speech, ridicule, caricature, or prejudice against specific social groups, including explicit references, stereotype reinforcement, or conventional assumptions about targeted groups.

Each comment is assigned multiple labels from ten categories, making the dataset multi-class and multi-labeled. Categories include *Civil* (devoid of hate speech), *Uncivil* (disparaging language or personal attacks), and various hate speech types targeting race/nationality, region, gender, religion, age, and sexual minorities.

One limitation of this dataset is the potential misclassification of neutral comments as hateful. For example, a benign statement referencing a group may be incorrectly flagged as hate speech. To mitigate this issue, we supplemented the dataset with additional neutral sentences following Kang et al. (2022).

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%). Figure 1 illustrates the label distribution.

4 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 exhibit significantly 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 et al., 2020).

The Gini index measures the overall dispersion of participation levels, reflecting how unequally comments are distributed among users. A higher Gini index indicates greater inequality in engagement. However, the Gini index has notable limitations in interpretation. Two distributions with identical Gini values can have different underlying structures, making it difficult to capture whether disparities are driven by the most or least active users. Additionally, the Gini index is more sensitive to changes in the middle of the distribution but less responsive to variations at the top and bottom.

To address these limitations, we incorporate the Palma index, which specifically measures the ratio of participation between the top 10% of commenters and the bottom 40%. An increasing Palma index indicates that the most active users are gaining even greater dominance over the least active users, highlighting the skewed nature of digital participation. Unlike the Gini index, the Palma index provides a clearer interpretation of who dominates the discourse in digital spaces and to what extent.

We applied these two metrics across different time periods, news domains, and news popularity rankings, depending on the specific analytical focus of each part of the study.

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:

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

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

To assess levels of comment hostility, we conducted a content analysis of comments from both heavy (top 10%) and light (bottom 40%) commenters, as defined by the Palma index. Within the top 10% group, we further distinguished the extreme top 1% from the remaining users, as a small subset of commenters appeared significantly more frequently than others.

As an initial step, we trained *KC-BERT*, a BERT-based deep-learning model (Lee, 2020), using the hate comment data described earlier. Following model training, we selected the best-performing version and applied it to a 1% stratified sample of comments from each engagement group. The re-trained model assigned a hostility score to each comment, and for simplicity, we categorized each comment based on its highest-scoring label while discarding those with all label scores below 0.5.

To facilitate analysis, we collapsed the ten predefined hate speech categories into three broader classifications: *civil*, *uncivil*, and *hateful*. Specifically, comments containing general profanity and personal attacks were classified as *uncivil*, while those with derogatory language targeting specific groups (e.g., race, gender, religion, region, etc.) were categorized as *hateful*. Comments devoid of such language were considered as *civil*. We then compared the distribution of comment types across user engagement groups, employing a chi-squared proportion test to determine whether differences in hostility levels between user groups were statistically significant.

5 Participation Inequality

Descriptive statistics on participation levels indicate a stark digital participation gap (Figure 3). On average, the top 10% of frequent commenters account for nearly half of all comments in news comment sections (50.11%), while the least active half (bottom 50%) contributes only 14.99% of total comments over the years. The figure clearly illustrates a consistent and substantial divide in digital participation, where a small subset of users disproportionately dominates the conversation. This imbalance underscores the motivation for our study, highlighting the need to investigate the structural disparities in online engagement.

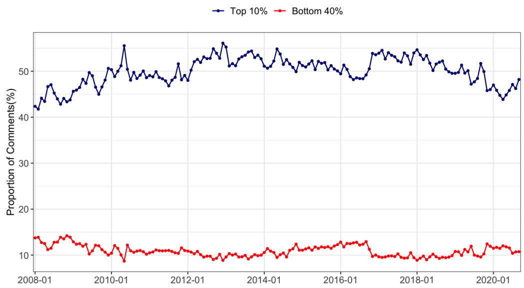

Figure 2: Share of Comments by Top 10% and Bottom 40% Gropus

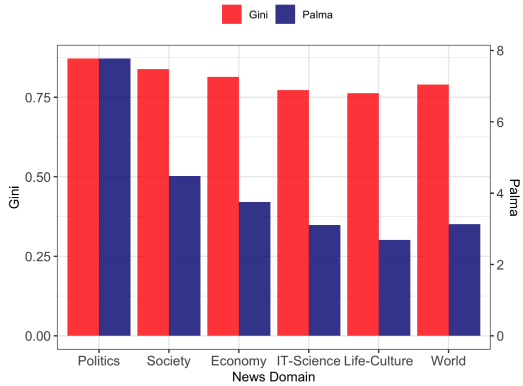

Figure 3: The Gini and Palma index Over Time by News Domain

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 indexes for different news stories based on their popularity ranks to assess how inequality changes across news interest levels.

Figure 3 illustrates participation inequality across different news domains, showing that political news exhibits the highest levels of inequality compared to other categories. Both Gini and Palma indexes reveal that Politics consistently stands out as the most unequal domain, indicating that discussions in political news sections are dominated by a small subset of highly active commenters. Society and Economy also exhibit relatively high participation inequality, though to a lesser extent than Politics. In contrast, domains such as Life/Culture and IT/Science display lower levels of inequality, suggesting that discussions in these categories are more evenly distributed among users.

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 indexes show a clear upward trend, indicating that as a news story becomes more popular, participation inequality increases. This pattern suggests that highly popular articles tend to be dominated by a small group of frequent commenters, while less popular articles see a more balanced distribution of participation. Among the different news domains, Politics and Society, again, consistently exhibit the highest levels of inequality across all levels of popularity, reinforcing the idea that digital participation gaps are most pronounced in politically charged discussions.

Taken together, these findings suggest that participation inequality is not only domain-specific but also influenced by news popularity. The more widely read an article is, the more concentrated the conversation becomes among a small subset of highly active users, particularly in Politics and Society.

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 RMD scores across different user groups, segmented into ten participation levels to capture finer distinctions beyond the broad bottom 40% and top 10% classifications. The figure shows that the least active commenter groups

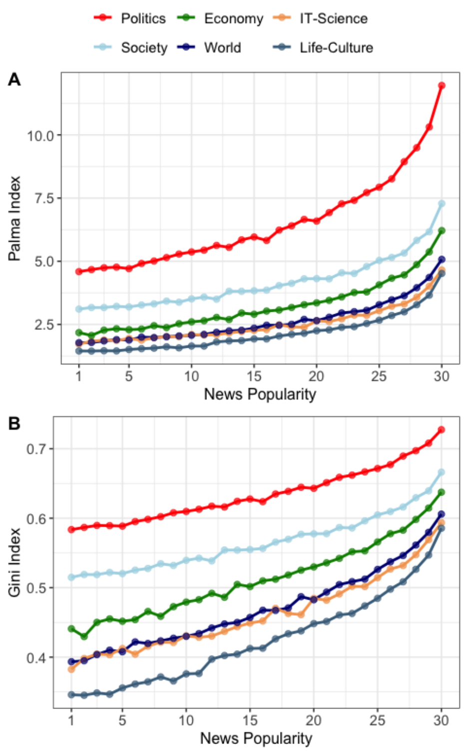

Figure 4: The Palma(Panel A) and Gini(Panel B) index by News Popularity

(Bottom 10% to Top 30-20%) cluster around zero, indicating that their participation closely aligns with the expected equal participation benchmark. In contrast, 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.

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 hyper-active commenters, reinforcing the severe imbalances in digital participation.

Participation Inequality and Political Events

Beyond these structural patterns, we now examine how participation inequality fluctuates in response to major political events, particularly during South Korea's electoral cycles and one of the most significant political events of the study period—the 2017 impeachment of President Park Geun-hye.

Figure 6 illustrates the Gini and Palma indices in the weeks leading up to three key political events: the 2012 and 2017 presidential elections and the 2016 impeachment of the president. The trends suggest that participation inequality intensifies as major political events approach, with both indices showing a marked increase in the final weeks leading up to each event. This pattern indicates that a small subset of highly active users becomes even more dominant in news comment sections during politically charged periods, further exacerbating the imbalance in online discourse. These findings suggest that political events act as catalysts for deepening participation inequality, amplifying the influence of highly engaged users while sidelining less active participants.

6 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 10% - < top 1%), and (3) the bottom 40% least active commenters. It is important to note that the top 1% and top 10% - < top 1% 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.

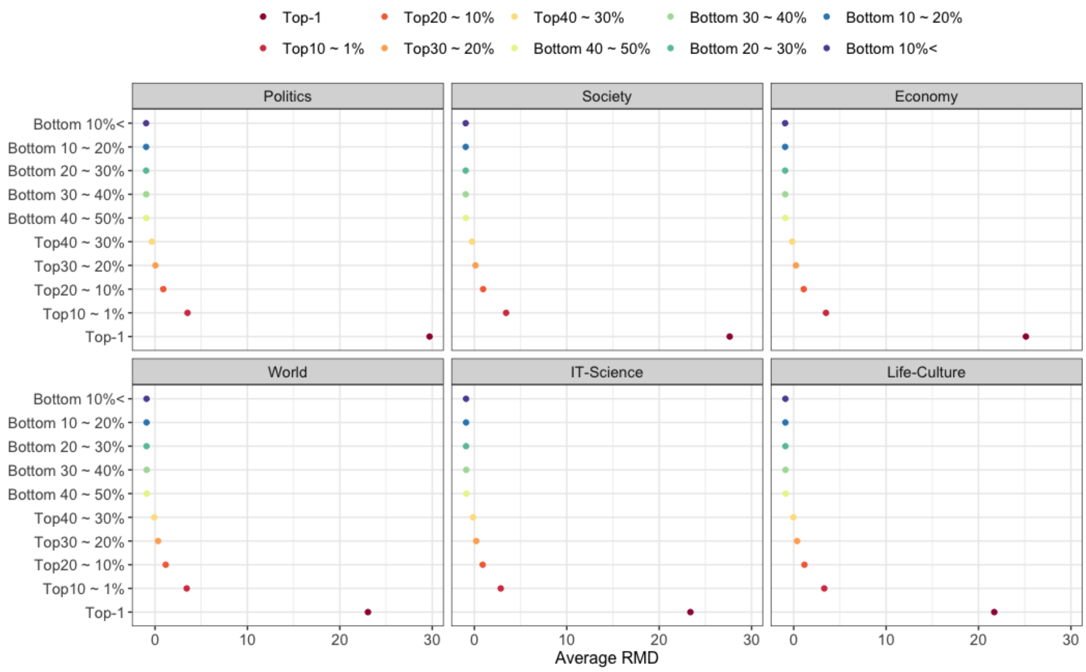

Figure 5: Average Relative Mean Deviation by News Domain

Figure 7 presents the distribution of comment categories across these three user groups. As expected, the most frequent commenters—the top 1% and top 10% - < top 1%—are significantly more likely to post *uncivil* comments compared to the less active bottom 40% (chi-square: 85.761, $p < 0.001$ for the comparison between bottom 40% and top 10% - < top 1%, and chi-square: 71.764, $p < 0.001$ for the comparison between bottom 40% and top 1%).

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 10% - < top 1% (chi-square: 139.19, $p < 0.001$). 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.

The disparity in hostility between active and inactive groups is still evident when examining differences across news domains. As shown in Figure 8, 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.

7 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, intensifying during major political events such as presidential elections. 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. This suggests that online discussions are not only dominated by a small fraction of users but are also skewed toward a more hostile or hateful discourse.

These findings carry important implications for digital public discourse and online platform governance. The dominance of a small, often hostile group in comment sections raises concerns about the representativeness of online discussions and their potential to skew public perceptions. Platforms aiming to foster healthier discourse may need to consider interventions that encourage broader participation while mitigating the outsized influ-

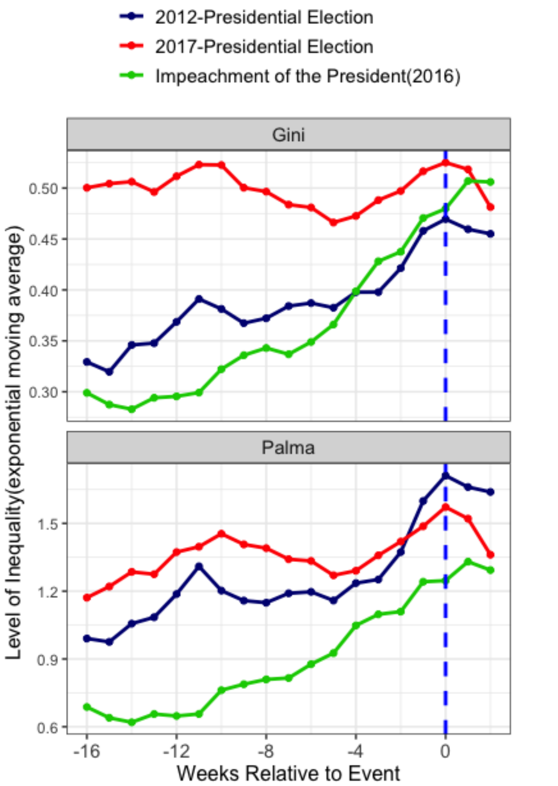

Figure 6: Participation Inequality Leading Up to Presidential Elections and the 2017 Impeachment

ence of highly engaged yet hostile users. Future research should further explore the causal mechanisms behind these dynamics and investigate potential strategies to counteract digital participation disparities and online hostility.

8 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

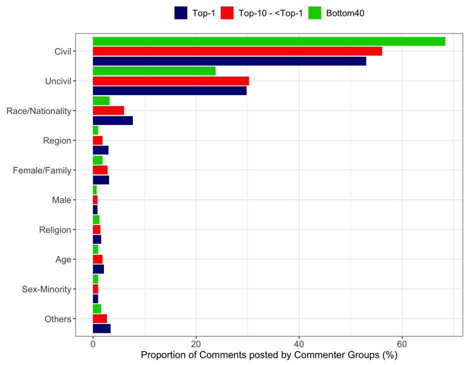

Figure 7: Hate Comment Classification Result by Percentile User Group

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, our study does not establish a direct causal relationship between participation inequality and online hostility. While our findings suggest that hostility is more prevalent among highly active users, we have not explicitly tested whether increasing inequality drives greater hostility or if other factors mediate this relationship. As participation inequality intensifies—especially during politically charged periods—aggressive discourse may become more concentrated among dominant commenters. However, 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 offers a foun-

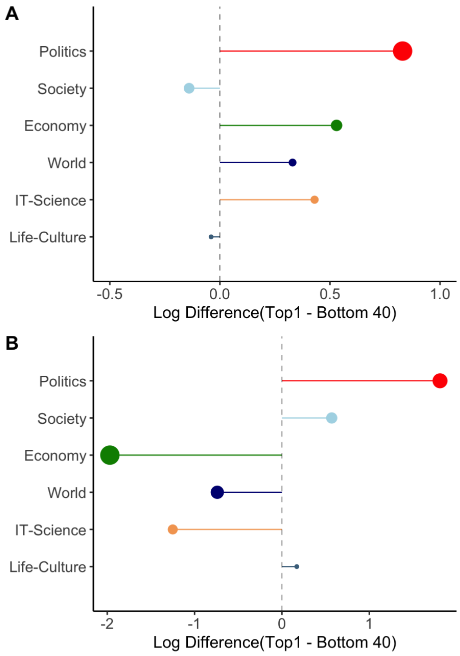

Figure 8: Log Difference in the Proportion of Uncivil (Panel A) and Hateful Comments (Panel B) between Extreme (Top 1%) and Inactive (Bottom 40%) User Group Across News Domains. Point sizes indicate the absolute difference in proportion.

dational analysis of how a vocal minority shapes digital discourse through both disproportionate engagement and increased hostility. Addressing these challenges in future research will be crucial for developing more effective moderation strategies and fostering healthier online discussions.

References

Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. 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*, pages 1035–1038, San Francisco USA. ACM.

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

Anees Baqir, Yijing Chen, Fernando Diaz-Diaz, Sercan Kiyak, Thomas Louf, Virginia Morini, Valentina Pansanella, Maddalena Torricelli, and Alessandro Galeazzi. 2023. Beyond Active Engagement: The Significance of Lurkers in a Polarized Twitter Debate. *Preprint*, arXiv:2306.17538. ArXiv:2306.17538 [physics].

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

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

Bradley Carron-Arthur, John A. Cunningham, and Kathleen M. Griffiths. 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.

Moumena Chaqfeh, Rohail Asim, Bedoor AlShebli, Muhammad Fareed Zaffar, Talal Rahwan, and Yasir Zaki. 2023. Towards a World Wide Web without digital inequality. *Proceedings of the National Academy of Sciences*, 120(3):e2212649120.

Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4):658–679.

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

Mattia Gasparini, Robert Clarisó, Marco Brambilla, and Jordi Cabot. 2020. Participation Inequality and the 90-9-1 Principle in Open Source. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–7, Virtual conference Spain. ACM.

Maria Glenski, Svitlana Volkova, and Srijan Kumar. 2020. User Engagement with Digital Deception. In Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, editors, *Disinformation, Misinformation, and Fake News in Social Media*, pages 39–61. Springer International Publishing, Cham. Series Title: Lecture Notes in Social Networks.

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

Eszter Hargittai and Aaron Shaw. 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.

Ariel Hasell and Brian E Weeks. 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.

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

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

TaeYoung Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh. 2022. Korean online hate speech dataset for multilabel classification: How can social science aid developing better hate speech dataset? *Preprint*, arXiv:2204.03262.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 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.

Seonghyun Kim. 2022. Korean unsmile dataset: Human-annotated multi-label korean hate speech dataset. https://github.com/smilegate-ai/korean_unsmile_dataset.

Youngjoo Kim, Yoonjin Shin, Hayoung Sim, Yoonjae Jang, and Park Mingyoo. 2021b. Media Users in Korea 2021. Technical report, Korea Press Foundation.

Vladimir Korovkin, Albert Park, and Evgeny Kaganer. 2023. Towards conceptualization and quantification of the digital divide. *Information, Communication & Society*, 26(11):2268–2303.

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

Gina M Masullo, Shuning Lu, and Deepa Fadnis. 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.

Jakob Nielsen. 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.

Se-Uk Oh, Ahran Park, and Jinho Choi. 2021. Digital news report in korea 2021. https://www.kpf.or.kr/front/research/selfDetail.do?seq=592216.

Mathilde Pak, Christophe André, and Jinwoan Beom. 2021. DIGITALIZATION IN KOREA: A PATH TO BETTER SHARED PROSPERITY? Technical report, Korea Economic Institute of America.

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

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

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

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

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

Anique Scheerder, Alexander Van Deursen, and Jan Van Dijk. 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.

Nicholas A Valentino, Ted Brader, Eric W Groenendyk, Krysha Gregorowicz, and Vincent L Hutchings. 2011. Election night's alright for fighting: The role of emotions in political participation. *The journal of politics*, 73(1):156–170.

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

Trevor Van Mierlo. 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.

Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. 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.

A Appendix

A.1 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 posting comments. This means we overlook the 90% of users, often referred to as "Lurkers" in the 90-9-1 principle.

Distribution of the Number of Comments

Online comment space is highly skewed. The histogram in Figure 10 indicates that the majority of users post one or two comments. When dealing with a highly skewed distribution, it is generally more appropriate to consider specific percentiles, as there is a significant difference in values between the top and the bottom of the distribution. Hence, this paper compares only top 10% and bottom 40% groups.

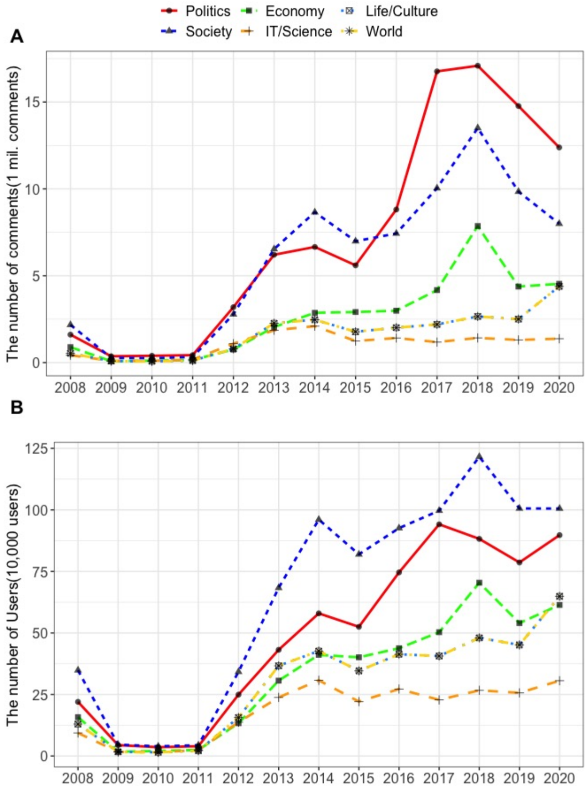

Figure 9: Change in the size of comment space: A. Change in the number of comments over time. B. Change in the number of users over time

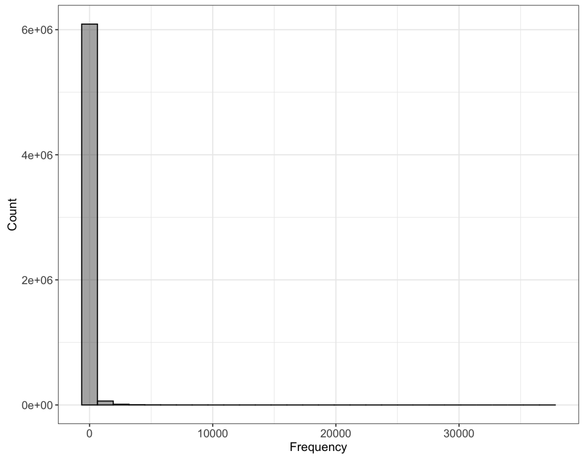

Figure 10: Histogram of the Comment Frequency

A.2 Training Performance(KC-BERT)

For training KC-BERT, we primarily trained two models: KC-BERT Base and KC-BERT Large. The KC-BERT Large model is larger than KC-BERT Base, with significantly more parameters.

To identify the model with the best performance, we conducted experiments using various hyperparameters, such as learning rate and batch size. Additionally, recognizing that the construction of the train/validation set could influence performance, we repeated the experiments with different configurations of the train/validation split.

The table 1 presents the best performance achieved by each model, with KC-BERT Base yielding slightly better results. The table 2 displays the classification performance for KC-BERT Base

Model	LRAP
KC-BERT base	**0.923**
KC-BERT large	0.92

Table 1: Model performance based on label ranking average precision(LRAP)

Category	Precision	Recall	F1-Score	Supprot
0	0.82	0.72	0.77	423
1	0.87	0.81	0.84	341
2	0.87	0.81	0.84	326
3	0.85	0.76	0.80	436
4	0.87	0.83	0.85	160
5	0.89	0.87	0.88	387
6	0.88	0.89	0.89	319
7	0.93	0.17	0.29	148
8	0.72	0.57	0.64	832
9	0.93	0.92	0.93	3990

Table 2: Classification Performance based on Precision, Recall, and F1-Score