



## The effect of online comment history disclosure on digital citizenship.

BY MINHYUNG LEE, JINYOUNG MIN, JUNYEONG LEE, CHANHEE KWAK, AND HANBYEOL STELLA CHOI

# A Small Clue Made of Fragmented Pieces

WHEN ONLINE INTERMEDIARIES provide a space for individuals to comment and share their opinions, the hope or expectation is to facilitate civic engagement. However, although civic engagement is achieved to a certain extent, uncivil behaviors, such as aggressively attacking opposing opinions, producing malicious comments and rumors, and even expressing insults and threatening comments, have also increased. Cyberbullying, cyber-sexual harassment, and online hate speech are only a few of the uncivil behaviors expressed via malicious comments. This behavior

is performed mostly in anonymous environments, where people are less concerned about the real-life impact of online behavior. According to an analysis of online comments, 53.3% of anonymous comments include hateful and aggressive language, while only 28.7% of real-name comments are uncivil.<sup>33</sup>

Anonymous interactions that inherently lack an understanding of their social context can lead to destructive behaviors, which could have extreme consequences. The negative implications of malicious online comments can be more serious for celebrities who must face many unknown individuals. For example, violent comments based on unconfirmed rumors have led to the deaths of K-pop stars.<sup>8</sup> Despite the victims' requests to refrain from making such comments, the production of offensive content continues behind the mask of anonymity. Uncivil comments are sometimes found in unexpected places. In response to an online article regarding miners trapped in a Chilean mine, the following comments were found: "These guys are frauds" and "We're just trying to make the world a better place one brainwashed, ignorant idiot at a time."<sup>39</sup>

As these cases demonstrate, un-

### » key insights

- The commenting function of online intermediaries and digital platforms is expected to facilitate civic engagement via healthy discussion and the sharing of diverse perspectives. However, while anonymous online spaces can encourage civil discussion, they can also nurture uncivil behaviors, including the posting of malicious comments.
- Several attempts have been made to resolve this issue, especially by decreasing anonymity; however, these solutions have often been less effective than expected or have had controversial consequences, such as limiting the freedom of speech.
- An online comment-history disclosure system provides users with access to their accumulated comment histories; helping them to perceive their digital identity is built upon their past behaviors can effectively induce digital citizenship behaviors while retaining their anonymity.



civil comments can degrade and skew public discourse and even further develop into violent threats. Even if these threats are not realized, abusive comments can cause enormous psychological harm, including fear, stress, depression, and feelings of inferiority.<sup>15</sup> Furthermore, malicious online comments can intimidate individuals who have already expressed their opinions, potentially discouraging them from participating in further discourse, thereby significantly limiting civic engagement.

Furthermore, the harmful effects of uncivil online behaviors are magnified since individuals tend to generate more aggressive comments in an anonymous online environment.<sup>31</sup> With the growing number of Internet users, the number of individuals posting malicious comments can easily become large enough to distort the direction and consequences of public discourse.

### How Anonymity Induces Uncivil Behavior

When people perceive anonymity online, they tend to be more willing to express their opinions, even on controversial and sensitive issues such as abortion.<sup>37</sup> The mere possibility of eliminating anonymity can decrease both the number of posts as well as the amount of dialogue that occurs in online communities.<sup>23</sup>

However, while anonymity enables individuals to share their opinions, it also allows them to express their opinions in an uncivil manner. For example, on *The Washington Post* website, which provides anonymity to commenters, uncivil online behavior is a significantly common occurrence.<sup>32</sup> The quality of comments on online articles decreases under anonymity and increases when the degree of anonymity decreases.<sup>30</sup> Individuals are more likely to post offensive comments when using non-social-media accounts rather than social-media accounts, which provide less anonymity.<sup>6</sup> This perceived anonymity significantly increases cyberbullying<sup>2</sup> and allows individuals to post malicious online comments because they feel less responsibility under anonymous conditions.<sup>22</sup>

How does anonymity, which makes online spaces rich environments for sharing opinions, also incite uncivil

behaviors? Social psychology answers this question using the deindividuation theory,<sup>40</sup> which suggests that an individual's unidentifiable deindividuated state in a crowd is the path to greater uninhibited expression. Individuals are also more likely to express negative emotions when they perceive a high level of anonymity, since negative emotions are less socially desirable than positive ones. Moreover, anonymous environments allow people to feel more comfortable in demonstrating less socially acceptable behaviors.<sup>35</sup> Under a state of anonymity, the burden of responsibility perceived by individuals when expressing their remarks, opinions, or certain behavior is lighter. This can lead them to extremely uncivil behaviors, such as bitterly criticizing the government or expressing hatred toward certain groups. Since the degree of perceived anonymity is related to community size,<sup>5</sup> online uncivil commenting behavior associated with anonymity will only worsen as the number of Internet users grows.

### Efforts to Deal with Uncivil Online Comments

Various attempts have been made to limit uncivil online comments. We identified three main methods for limiting uncivil comments: human control, algorithmic control, and environmental control.

**Human control.** The human-control method does not aim to control the anonymity embedded within online environments but rather seeks to use peoplepower to directly deal with malicious comments. A representative attempt to implement this technique is the use of professional moderators. Online sites using this method encourage their users to flag or report offensive comments; subsequently, professional moderators scrutinize these comments to determine whether they must be removed. However, users tend to use flags not only to report offensive comments but also as individual tactics to express disagreement, retribution, and harassment,<sup>13</sup> which further increases the difficulties associated with moderating comments. Another technique, *interactive moderation*, goes beyond simple moderation; it uses replies and counter-speech to

deal with malicious comments. Although this method can facilitate a deliberative atmosphere that encourages user participation, counter-speech in response to hate speech has the side effect of giving hateful comments more exposure.<sup>11</sup> Additionally, counter-comments made by community moderators are not as effective as those made by other users,<sup>26</sup> suggesting that other commenters, rather than moderators, play an important role in preventing uncivil behavior. Therefore, the interactive moderation method is not as effective as expected, especially considering the significant costs associated with hiring and running moderators.

Another human-control method is limiting entire comment sections, which is relatively simpler than professional moderation. This is classified as a human-control method because it does not control the anonymous environment per se but requires human intervention to determine whether an article or article category is suitable for commenting. For example, *The New York Times* and CNN have an optional commenting system that restricts the articles that can be commented on. Although eliminating the comment section is a possible solution for preventing uncivil comments, it also eliminates the entire possibility of healthy online civic engagement.

**Algorithmic control.** This method depends heavily on technology to filter and delete malicious comments; however, it does not deal with the anonymity embedded in this context. Since this method is used by online news organizations to filter and remove hate speech, it often encounters difficulties in defining hate speech<sup>11</sup> and is much criticized as being a "discriminating gatekeeper."<sup>18</sup> Therefore, most methods for controlling uncivil comments use algorithmic control as a complement. For example, blacklisting or blocklisting combines human- and algorithmic-control methods, enabling users to avoid viewing malicious comments by blocking certain users' comments. X (formerly Twitter) uses blocklisting applications such as Block Bot and Block Together. However, making the most efficient and effective but unbiased lists may be a challenging task, since blocking a particular user's account hides not only their individual



comments but also all the comments ever made by that account, and subscribing to existing blocklists means blocking an entire group of accounts on the list. For example, managing community-curated blocklists faces difficulties in motivating and coordinating community moderators.<sup>19</sup> Algorithmically curated blocklists, which use predefined criteria to make lists, are not only ineffective in making complex curation decisions<sup>19</sup> but are also criticized as being biased and discriminatory.<sup>18</sup> Additionally, the algorithmic-control method focuses on comment readers, providing them with the option of selecting what should not be shown to them rather than preventing commenters from posting malicious comments. More importantly, blocklisted accounts do not face any constraints in terms of posting further malicious comments.<sup>19</sup> Therefore, using blocklisting as the sole method to control malicious comments has severe limitations.

**Environmental control.** Another method of controlling malicious commenting behavior is environmental control, which attributes the source of malicious behavior to the anonymity of the commenting environment. While the human- and algorithmic-control methods seek to directly deal with malicious comments or accounts, environmental controls create signals or an atmosphere wherein malicious commenting activities eventually lead to certain consequences that are directly attributed to a particular commenter. Therefore, environmental controls mostly involve the elimination of anonymity.

Since individuals tend to demonstrate more trust, accountability, and intention to cooperate with others in a non-anonymous environment,<sup>25</sup> when a speaker is identifiable, the number of positive sentiments tends to increase, while the number of negative expressions decreases during discussion.<sup>16</sup> Therefore, to solve the problem of malicious comments by directly increasing identifiability and thus decreasing anonymity, an attempt to introduce a real-name Internet system has emerged and has been intensively discussed. While this system is associated with several controversies related to free-speech and public-discourse restric-



**While anonymity enables individuals to share their opinions, it also allows them to express their opinions in an uncivil manner.**



tions, the right of self-determination in terms of disclosing personal information, and the possibility of monitoring citizens, the world's first real-name Internet system was legislated in Korea in 2006. The law mandated that portal sites with more than 0.3 million daily users and online news bulletins with over 0.2 million daily users must implement the system. Thirty-seven sites became the subjects of this legislation, including Naver and Daum, the largest portal sites in Korea. Subsequently, users appeared to be more careful when leaving comments, because it was easy to determine who and what someone had posted since their real names were exposed. Under the real-name system, the number of malicious comments decreased. However, the total number of comments also decreased, thus limiting public discourse. The system was constantly mired in controversy. Eventually, a constitutional petition was submitted in Korea. In 2012, the law was ruled unconstitutional and abolished, with lawmakers saying it undermined the freedom of anonymous expression, the intermediaries' right to provide freedom of speech, and the users' right of self-determination in disclosing personal information.


In contrast, in the U.S., enforcing a real-name comment system has never been an option because it goes against the First Amendment, which values freedom of speech. However, in the wake of “The Great De-platforming” and the spread of misinformation about COVID-19, voices are being raised for legislation to control anonymity to decrease online abuse, hate speech, and racism. Thus, policymakers are considering a legal solution to reduce the anonymity of online users—for example, by revising Section 230. Even individuals who oppose this legislation acknowledge the need to control anonymity. They suggest that online intermediaries or digital platforms—private actors not ruled by the First Amendment—should employ their own methods for controlling anonymity, such as authentication or stratified identification.

In summary, the human- and algorithmic control methods do not prevent the posting of malicious comments, as they are mostly post hoc methods that address malicious comments only af-


ter they have been posted. In contrast, the environmental-control method, aimed at controlling the anonymity of online environments, seeks to prevent the posting of malicious comments. This method views anonymity, which was considered to be an absolute virtue, as something between an inevitable feature of Internet design and a controllable feature of architectural design. However, eliminating or severely limiting anonymity is viewed as either unconstitutional or doing more harm than good. Therefore, there is a need for a different approach that does not aim to limit anonymity but rather produces an effect like those induced by decreasing anonymity.

#### **A Different Approach: Publication of Online Comment History to Build a Sense of Digital Identity**

An essential characteristic of anonymity is the inability to track individuals by their online behavior since it is not linked to their respective identities. Another important aspect of online anonymity is that anonymous online behavior is generally performed under individually different anonymities—that is, anonymity not only detaches a behavior from its agent but also detaches one behavior from another. In other words, while the same person may create several comments, there is no way to identify, collect, or classify the comments by their agent. Consequently, there is no sense of identity. Focusing on individually different anonymities and creating a sense of digital identity, as opposed to whether a real person can be identified or not, a different attempt stemmed from the idea that if an individual's online comments are accumulated under a unique pseudo-identifier such as a nickname, there will be a sense of digital identity built from them, even in an anonymous environment. Although a person's digital identity is not linked with a person's real identity, it will make the person perceive that their online comments are now only detached from their offline identity but are still attached to their digital identity. Once it is formed, regardless of the personas, digital identity cannot be completely ignored or forgotten and can be subject to impression management and self-presentation.



**This new approach is expected to limit the perceived anonymity and increase the sense of identity without impairing the anonymity of the Internet.**



Therefore, this new approach is expected to limit the perceived anonymity and increase the sense of identity without impairing the anonymity of the Internet. In South Korea, the necessity to control anonymity originates from the field of online news consumption. News portal sites in Korea offer articles and videos produced by various news channels in a mash-up format; comment sections allow users to share their opinions with others. Among these, Naver is the largest portal in South Korea, with more than 30 million monthly active users. About 68.6% of Korean adults use Naver as their primary news consumption channel.<sup>21</sup> Naver users read the news and comments on the news posted by others to get a sense of how the public views the issues.<sup>34</sup> Owing to its impact, there are countless incidents of opinion manipulation and comment fabrication on Naver.

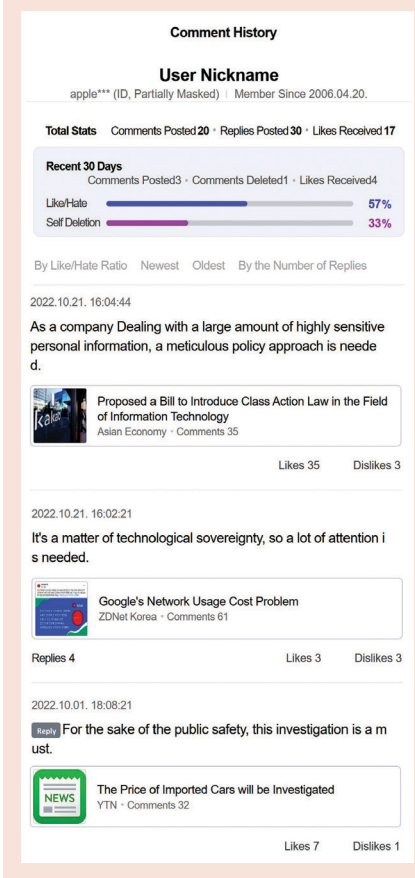
To alleviate the social problems caused by malicious comments, such as cyberbullying, Naver decided to disclose the online comment histories of all its users starting March 19, 2020. This decision was announced to the public a day before its implementation. Therefore, users did not have sufficient time to find and delete any comments they did not want to show the public. After that date, users were able to view the comment history of other users on the Naver news website. This new system discloses a user's comment history, including their screened ID; nickname; list of comments; the news articles on which they were posted; reactions of other users to recent comments, such as likes and dislikes; and comment-management measures, such as the number of deleted comments.

This implementation can be viewed as a method that executes the concept of accountable anonymity. Accountable anonymity does not limit users' anonymity as long as they communicate legitimately. However, when something prohibited occurs, the accountable anonymity system collects evidence of the wrongful behaviors of abusive users and establishes a link between their activities and their identity without affecting the anonymity of other users. This accountable-anonymity system requires some as-

pects of a user's identity so they can be traced when necessary—for example, from their connection (IP) address at the minimum level to their real identity at its full scale. These can be collected by service providers or trusted third parties (for example, authentication centers).<sup>1</sup> Then, it uses certain cryptographic mechanisms<sup>12</sup> to establish a certain level of anonymity, such as selectively deanonymizing, pseudonymizing, and maintaining the anonymity of identity information, depending on the various types of identity information.<sup>36</sup> Therefore, since users are aware that while their anonymity can be maintained, their identity can be revealed when they participate in abusive behaviors, accountable anonymity is expected to serve as a deterrent to the abuse of anonymity, which can lead to uncivil behavior.<sup>38</sup> Thus, the concept of accountable anonymity is an attempt to reduce the adverse effects of anonymity while maintaining its merits. However, an online comment history disclosure system is not simply about building a link between users' activities and their real identities. Since Naver requires the verification of a mobile phone number when creating a user account (up to three Naver accounts can be created using one mobile number), users are not entirely free from the possibility of being identified, which means that the concept of accountable anonymity is always in action at some level.<sup>a</sup> However, Naver still had to implement several policies to decrease malicious comments before introducing its online comment-history disclosure system. For example, the number of “upvote/downvote” clicks by a single account within 24 hours was limited to 50, and an option to sort online news

a To create a Naver account, an individual's mobile phone number is required, and in Korea, mobile phone numbers are only issued to individuals who have completed an identity verification process. The process of identity verification when creating a Naver account is done in a way such that Naver only obtains the result of whether a particular user actually owns a mobile phone number from a telecommunication company. In other words, any identity information that can be used to identify a person is not stored by Naver. In terms of accountable anonymity, in this case, the telecommunication company becomes a third-party company that can restore an individual's identity if necessary.

**Figure 1. Example of online comment history (translated from Korean to English).**



comments in the political news section was adjusted from most liked to newest-first only. Moreover, Naver provided each news company with the discretion as to whether to provide a news comments section for a particular article, whether to display news comments below the text, and how to sort news comments in the Naver news section. Additionally, the comment sections under all news items were hidden by default and only shown upon an explicit click. Furthermore, the news comment function was entirely removed to prevent the generation of malicious comments under entertainment news. Therefore, online comment-history disclosure does not just decrease anonymity through accountable anonymity but also provides a clear sense of the decreased anonymity established by building a digital identity.

Moreover, online comment-history disclosure systems can be viewed as similar to reputation-based systems, such as Airbnb and Uber, wherein user reputations are evaluated based on accumulated reviews.<sup>17</sup> However, such

two-sided reputation systems differ from online comment-history disclosure systems in that the reputations of each side are explicitly based on the other side of the transaction in the former's case, while only the commenters' reputations are built indirectly based on their own comments in the latter's case. In online comment disclosure systems, a commenter's previous comment records can be used to determine the user's tendencies. Figure 1 presents Naver's comment history page. Once an individual posts comments on a news page, the news page and the individual's comments are automatically collected within the commenter's account. When the reader of a comment clicks on the screened ID provided at the top of a comment, a page (Figure 1) displays information about the commenter's comment history, including statistics and the contents of previous comments.

### Analysis of the Effects of the Publication of Online Comment History

Korean news media companies distribute their content through various channels, which include not only their own websites but also large portal sites. An interesting trend in news formats is that Koreans prefer video-type news rather than text-type articles,<sup>21</sup> and video platforms such as YouTube have emerged as a new channel for news consumption. In response to this trend, news media companies have started using YouTube as their regular channel for uploading news. This means that the exact same content can be found on both YouTube and Naver. These two outlets are similar in that their users can consume news content and share their opinions without any restrictions. However, Naver's decision to disclose users' comment histories may have created a different environment for comment writers. Compared with YouTube's lack of access to user-comment records, Naver's provision of access to users' accumulated and trackable comment histories may significantly affect their comment writing behavior. Therefore, to determine the effects of an online intermediary using an online comment-history disclosure system, we compare this use with an intermediary that does not implement

such a system. In this study, we analyzed the comments on news publications available on Naver and compared them with those on YouTube.

**Video news selection.** Naver presents video news produced by Korean news media companies; therefore, people can watch video news similar to how they watch on YouTube. Moreover, individuals can post comments for video news on both platforms. We collected the digital comments for Korean video news available on both platforms between November 2019 and June 2020. The video news whose comments we collected is listed identically on both platforms, thereby enabling us to compare user comments to ensure any differences are not caused by variation in the content. We randomly selected news from diverse categories to avoid any biases that may arise from category characteristics. These included politics (51%), society (24%), world (19%), business (4%), lifestyle (1%), IT (0.004%), and others (1%). News articles in the form of video news were obtained from various broadcasting companies, namely YTN (29%), MBC (19%), JTBC (16%), Channel A (9%), KBS (8%), SBS (8%), TV Chosun (7%), and MBN (4%). Consequently, we collected 6,398 video news items. Since we intended to analyze the user comments on these items, we only included news that had comments on both platforms. Therefore, we used 6,262 video news items (3,131 on each platform) for further analysis. The total number of comments on these 6,262 video news items was 1,722,343 (954,200 for Naver and 768,143 for YouTube), while the average number of comments was 275.05 (304.76 for Naver and 245.33 for YouTube).

**News comment sentiment analysis.** We performed sentiment analysis (SA) on news comments to identify the uncivil/civil behaviors of commenters. In machine learning (ML), SA and related approaches have been widely used, especially for natural language processing (NLP) and information retrieval. These approaches have been facilitated by the increased availability of large datasets and the development of commercial intelligence applications.<sup>14,28</sup> To calculate the sentiment scores of the comments on the selected video news, we applied an automated SA

technique in our analysis. SA computationally detects emotion, opinion, sentiment, and subjectivity in text.<sup>24,29</sup> By successfully mining opinions and measuring emotions, SA helps to accomplish two tasks: detecting the sentiment signals of text segments (for example, sentences) and measuring the strength and polarity of the sentiment within those segments.<sup>29</sup> To perform this SA for Korean data, we employed the rhinoMorph morphology analyzer, whose dictionaries are based on the Korean Modern Tagged Corpus, which includes a 12-million-phrase scale, created by the Korean government.<sup>9</sup> After model training, we applied the GridSearchCV method, which sequentially inputs hyperparameters into the model to verify the generalization error and determine the optimal parameters for the trained model. Using these methods, we measured the sentiment of each comment in the selected video news. Sentiment value is 1 when a comment is classified as a positive comment and 0 if not. Next, we calculated the percentage of positive comments for each video news item.

**Analysis of the impact of comment-history disclosure.** To validate the causal relationship, we employed the difference-in-differences (DID) model, which is widely used to estimate causal relationships.<sup>3</sup> This model was appropriate for our analysis since it provides an experimental design with observational data because the treatment (that is, implementation of an online comment-history disclosure system) occurred on one of the platforms during the observation period. This allows us to compare the treatment group (comments on Naver) with the control group (comments on YouTube) and observe the changes that occur within each group before and after implementation of the comment-history disclosure system. The following equation represents our model:

$$\begin{aligned} \ln(\text{AvgSent}_{ijt}) &= \beta \times (\text{Treatment}_{ij} \times \\ &\quad \text{AfterImplementation}_t) \\ &\quad + \alpha \times \text{Treatment}_{ij} \\ &\quad + \gamma \times \text{AfterImplementation}_t \\ &\quad + X_{ijt} + \delta_i + \varepsilon_{ijt} \end{aligned}$$

In the above equation,  $\text{AvgSent}_{ijt}$  refers to the average sentiment value

of all comments regarding video news  $i$  listed on platform  $j$  at time  $t$ .  $\text{Treatment}_{ij}$  is a binary variable that indicates whether video news  $i$  is displayed on a platform  $j$  that implements a comment-history disclosure system (Naver) or does not (YouTube).  $\text{AfterImplementation}_t$  refers to whether the time  $t$  occurs after the implementation of a comment-history disclosure system. In this study, the disclosure system was implemented in March 2020; thus, the variable  $\text{AfterImplementation}_t$  takes the value of 1 if the data is after March 2020 (that is, from April 2020 to June 2020) and 0 if it is before March 2020 (that is, from November 2019 to January 2020). Since Naver often deploys policies to create a less harmful commenting environment, we selected the period in which the effects of the implemented comment-history disclosure system can be captured. This period provides consistent data for all the variables in our model and covers the periods before and after policy implementation. The variable of our interest,  $\text{Treatment}_{ij} \times \text{AfterImplementation}_t$ , represents the effects of implementing a comment-history disclosure system.  $X_{ijt}$  denotes the related control variables, such as the average number of likes in the replies to comments for video news  $i$  listed on platform  $j$  at time  $t$  ( $\text{AvgLikesReply}_{ijt}$ ), the average number of likes in the comments for video news  $i$  listed on platform  $j$  at time  $t$  ( $\text{AvgLikes}_{ijt}$ ), and the average number of hates in the comments for video news  $i$  listed on platform  $j$  at time  $t$  ( $\text{AvgHates}_{ijt}$ ).  $\delta_i$  includes unobserved video news-specific characteristics as well as observed characteristics, such as the news section category to which video news  $i$  belongs to ( $\text{Section\_IT}_i$  for the IT section,  $\text{Section\_Biz}_i$  for the business section,  $\text{Section\_Lifestyle}_i$  for the lifestyle section,  $\text{Section\_World}_i$  for the world section, and  $\text{Section\_Politics}_i$  for the politics section) and the news media company that published the video news ( $\text{BC\_KBS}_i$  when video news  $i$  is published by KBS;  $\text{BC\_MBC}_i$  when published by MBC;  $\text{BC\_SBS}_i$  when published by SBS,  $\text{BC\_TVC}_i$  when published by TV Chosun,  $\text{BC\_YTN}_i$  when published by YTN,  $\text{BC\_ChA}_i$  when published by Channel



A, and  $BC\_JTBC_i$  when published by JTBC). These are binary variables with a value of 1 if yes and 0 otherwise.

### Findings of the Difference-in-Differences Analysis

Table 1 summarizes the results of the DID model. In Model 1, we only included the DID variables, while we included control variables in addition to the DID variables in Model 2. For all models, the coefficients of the variables of interest ( $Treatment_{ij} \times AfterImplementation_t$ ) are significantly positive, indicating that the sentiment of the comments in news articles is more positive after implementing a comment-history disclosure system. Specifically, the average sentiment value of all comments for video news  $i$  in Naver, which implemented an online comment-history disclosure system, increased by 24.1% for Model 1 and 24.5% for Model 2 after the system was implemented, compared with YouTube, which does not implement such a system. Among the control variables, the average number of likes in the replies to comments ( $AvgLikesReply_{ijt}$ ) also has a significant impact on the average sentiment value of the comments. Model 2 shows that the average sentiment value of all comments on video news increases by 0.7% as the average number of likes increases by one unit.

According to the DID estimation results presented in Table 1, the implementation of a comment-history disclosure system decreases the number of negative comments. Figure 2 depicts the main results, showing how the average sentiment in article comments changes with the implementation of a comment-history disclosure system and group differences in the changes in average sentiment. The average sentiment of comments on the video news became more positive after implementing the system compared with that of the comments before implementation. Since individuals perceive that their digital identity is built upon their past behavior and the commenting environment is not entirely anonymous, they may feel more responsible for their behavior and opinions and may refrain from indiscriminate criticism.

**Additional analysis considering sentiment measure.** Assessing the

**Table 1. DID estimation results.**

$\ln(AvgSent_{ijt})$	Model 1	Model 2
$Treatment_{ij}^*$	0.241*** (0.007)	0.245*** (0.008)
$AfterImplementation_t$	—	—
$Treatment_{ij}$	-0.251*** (0.006)	-0.275*** (0.008)
$AfterImplementation_t$	-0.584*** (0.169)	-0.619 (0.158)
$AvgLikesReply_{ijt}$	—	0.007*** (0.001)
$AvgLikes_{ijt}$	—	0.001 (0.001)
$AvgHates_{ijt}$	—	0.001 (0.004)
Video-News Fixed Effect	Included	Included
Day Fixed Effect	Included	Included
Number of Observations	6,262	6,262
R-Squared	0.757	0.761

Robust standard errors in parentheses.  
\* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

negativity of online content is the most conventional and effective way to identify malicious comments. However, it is difficult to assess the maliciousness of an online comment based solely on the negativity measure.<sup>4</sup> Therefore, we employed the hatred ratio of each comment by applying a recently developed multi-label Korean online hate-speech algorithm.<sup>20</sup> To classify comments as hate speech, the algorithm combined the findings of existing studies<sup>20</sup> and considered the main issues that Koreans pay attention to. Consequently, the algorithm classified Korean perceived hate speech into eight categories: non-hate speech and whether a com-

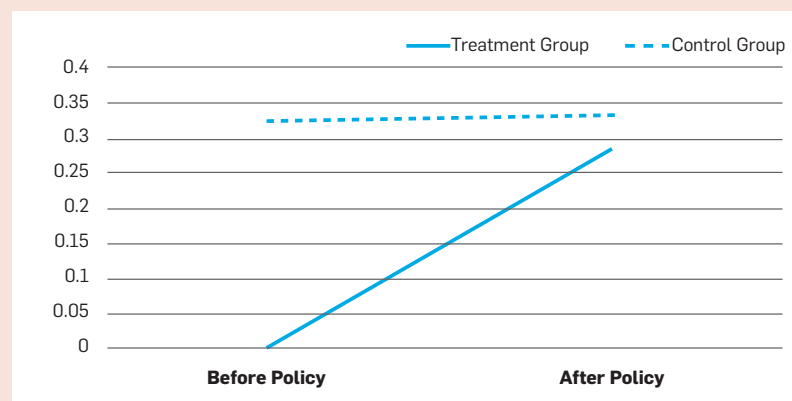
ment expresses hatred toward either female-, male-, queer-, generation-, religion-, race-, or region-related issues. This algorithm was further developed by training 24,000 online news comments collected from major Korean Web portals, such as Naver and Daum, based on three multi-label classifiers: KcBERT-base, KcBERT-large, and KcELECTRA-base. Using this algorithm, we measured the degree of each hate category for each online comment in our data.<sup>b</sup> We obtained the average value of this measure by aggregating online comments by news  $i$  on platform  $j$  at time  $t$ . By employing the new variables as the dependent variables, we analyzed how a comment-history disclosure system affects the generation of hate comments. Table 2 depicts the additional analysis results. According to the analysis, we find that the average hatred ratio of each category of the online news comments declined after system implementation.

### Discussion and Conclusion

While online discourse can help to achieve deliberative democracy via civic engagement that enriches diverse perspectives and promotes healthy discussion, uncivil comments hinder this opportunity by distorting the direction of online discourse and creating unnecessary conflict. To ensure that everyone can express their opinions without the fear of being shunned owing to anonymity, it is im-

<sup>b</sup> The analysis result of racial hate is not reported because the presence of racial hate in Korean online comments was low and the effect of racial hate was not significant.

**Figure 2. Average change in sentiment before and after policy implementation.**





**Table 2. Additional analysis considering the Sentiment Measure.**

	$\ln(\text{NoHate}_{ijt})$	$\ln(\text{FemaleHate}_{ijt})$	$\ln(\text{MaleHate}_{ijt})$	$\ln(\text{QueerHate}_{ijt})$	$\ln(\text{GenerationHate}_{ijt})$	$\ln(\text{ReligiousHate}_{ijt})$	$\ln(\text{RegionalHate}_{ijt})$
Treatment <sub>ij</sub> * AfterImplementation <sub>t</sub>	0.062*** (0.005)	-0.137*** (0.018)	-0.135*** (0.013)	-0.091*** (0.012)	-0.332*** (0.024)	-0.142*** (0.018)	-0.195*** (0.027)
Treatment <sub>ij</sub>	-0.144*** (0.005)	0.328*** (0.018)	0.220*** (0.013)	0.215*** (0.012)	0.647*** (0.025)	0.346*** (0.018)	0.544*** (0.028)
AfterImplementation <sub>t</sub>	-0.183* (0.106)	-0.631 (0.460)	0.180 (0.280)	-0.004 (0.257)	-1.338*** (0.348)	0.142 (0.169)	0.845 (0.520)
AvgLikesReply <sub>ijt</sub>	-0.001 (0.001)	0.010*** (0.003)	0.007*** (0.002)	0.004** (0.002)	0.001 (0.002)	0.004** (0.002)	-0.001 (0.003)
AvgLikes <sub>ijt</sub>	-0.001 (0.001)	-0.011** (0.004)	-0.003 (0.002)	-0.001 (0.002)	0.001 (0.005)	0.003 (0.003)	-0.002 (0.006)
AvgHates <sub>ijt</sub>	0.002 (0.003)	0.007 (0.011)	0.009 (0.008)	0.013** (0.006)	-0.030** (0.012)	-0.032*** (0.011)	0.016 (0.016)
Video-News Fixed Effect	Included	Included	Included	Included	Included	Included	Included
Day Fixed Effect	Included	Included	Included	Included	Included	Included	Included
Constant	-0.070 (0.105)	-4.117*** (0.460)	-5.692*** (0.279)	-5.207*** (0.250)	-3.714*** (0.348)	-5.487*** (0.167)	-5.492*** (0.465)
Number of Observations	6,262	6,262	6,262	6,262	6,262	6,262	6,262
R-Squared	0.837	0.824	0.741	0.860	0.729	0.861	0.781

Robust standard errors in parentheses. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

portant to foster an atmosphere that values digital citizenship, which suggests how to maintain our integrity and dignity in our interactions with other people as social beings. Therefore, what we should ultimately concern ourselves with is how we nurture digital citizenship, so healthy civic engagement is naturally nourished.

Digital citizenship is built upon one's responsible behavior in online environments. In this sense, although online intermediaries have undertaken various approaches to address uncivil comments, they have experienced varying amounts of success in terms of nurturing digital citizenship. For example, in a system that uses professional moderators, the final responsibility and discretion regarding the retention and deletion of comments lies with the moderators. Although individual users can contribute to this system by flagging and reporting malicious comments, the existence of final decision makers makes users

less responsible for creating a healthier commenting environment.<sup>27</sup> Moreover, blacklisting and blocklisting do not place the burden of building a less harmful commenting environment on individual users either. For most users, blacklisting and blocklisting are not about responsibility but more about using their own discretion in terms of deciding whether to view malicious comments. Algorithmic control methods are similar in that they do not involve individual users taking responsibility for their behavior, leaving technology in charge and users in a position to merely apply what the technology provides to them.


In contrast, environmental-control methods emphasize the responsible behavior of users and even their accountability. In this case, individual users may carefully consider their own commenting behavior. However, simply eliminating anonymity and using a real-name system does not increase digital citizenship. For

example, Facebook has used a real-name system since it began as a social networking service to link offline relationships. Today, it plays the role of a digital intermediary whose space nurtures public discourse, including that of various hate groups. While many intermediaries undertake significant efforts to remove and filter hateful content, Facebook is known to take less responsibility for such content, adopting a relatively hands-off approach, which ironically makes a case for users to take responsibility for their content since they use their real names. This is partly because Facebook's design tends to separate users into homogenous groups, such that certain extremist or hate speech may be magnified and reinforced within these groups.<sup>10</sup> This example indicates that simply introducing the concept of controlling anonymity might not work in reality if it is not accompanied by careful technological design and implementation. This is evident since

online anonymity regulations cannot be executed, do not reduce, and even increase online malicious commenting.<sup>7</sup> Since technologies can help to tailor user behavior and control how online spaces are deployed, the role of such intermediaries is crucial.

Overall, as previous attempts indicate, placing constraints on commenting is ineffective and often perceived as limiting free speech. The filtering of uncivil comments is not as effective a method as expected. The elimination of anonymity, the main source of uncivil commenting behavior, not only limits individuals' right to free speech but is also not particularly effective when its implementation is not carefully designed. In this sense, a more effective solution may be an environmental control that does not limit user anonymity, facilitates deliberative discourse, and provides an indirect and subtle nudge that builds a sense of identity among users. After all, shaping the behavioral norms of users by tailoring the way users see, feel, and experience technology is what digital platforms do best. Therefore, we argue that environment controls, along with careful architectural design, implemented to decrease perceived anonymity can be effective, but only if it is designed to nurture digital citizenship, thereby urging online intermediaries to thoughtfully and carefully shape and influence user behavior. We are not suggesting that an online comment-history disclosure system is the only method to establish a less harmful commenting environment; multiple methods can be combined and implemented together. Technological development and thoughtful thinking are essential to developing new approaches to achieve safer commenting environments. Regardless of the methods used, emphasizing the users' role in creating a healthy online environment is crucial when designing online platforms. Our analysis of the implementation of an online comment-history disclosure system that helps to build a digital identity shows the possible ways in which online intermediaries and digital platforms can facilitate digital citizenship, further emphasizing the importance of their roles.

## Acknowledgment

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A2A03041910). Jinyoung Min is the corresponding author. 

## References

- Ahmad, W. and Clark, D.D. A systems approach toward addressing anonymous abuses: Technical and policy considerations. *IEEE Security & Privacy* 19, 2 (2021), 38–47.
- Bartlett, C.P. et al. Predicting cyberbullying from anonymity. *Psychology of Popular Media Culture* 5, 2 (2016), 171.
- Bertrand, M. et al. How much should we trust differences-in-differences estimates? *The Quarterly J. of Economics* 119, 1 (2004), 249–275.
- Bridges, J. and Vásquez, C. If nearly all Airbnb reviews are positive, does that make them meaningless? *Current Issues in Tourism* 21, 18 (2018), 2057–2075.
- Bunn, J.Y. et al. Urban-rural differences in motivation to control prejudice toward people with HIV/AIDS: The impact of perceived identifiability in the community. *The J. of Rural Health* 24, 3 (2008), 285–291.
- Cho, D. and Acquisti, A. The more social cues, the less trolling? An empirical study of online commenting behavior. In *Proceedings of the Workshop on the Economics of Information Security* (2013).
- Cho, D. and Kwon, K.H. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior* 51 (2015), 363–372.
- Cho, J. Deaths of Goo Hara and Sulli highlight tremendous pressures of K-pop stardom. *ABC News* (2019).
- Choi, S. Implementation of Open Type Korean morphological analyzer based on collective intelligence. *Language & Information Society* 22 (2014).
- Cinelli, M. et al. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.
- Citron, D.K. and Norton, H. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review* 91, (2011), 1435.
- Claessens, J. et al. Revocable anonymous access to the Internet?. *Internet Research* 13, 4 (2003), 242–258.
- Crawford, K. and Gillespie, T.J.N.M. and Society What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
- Das, S.R. and Chen, M.Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53, 9 (2007), 1375–1388.
- Delgado, R. and Yun, D. Neoconservative case against hate-speech regulation—Lively, D'Souza, Gates, Carter, and the Toughlove crowd. *Vanderbilt Law Rev.* 47, (1994), 1807.
- Festinger, L., Pepitone, A., and Newcomb, T. Some consequences of de-individuation in a group. *J. of Abnormal and Social Psychology* 47, 2 (1952), 382–389.
- Fradkin, A., Grewal, E., and Holtz, D.J.M.S. Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb. *Marketing Science* 40, 6 (2021), 1013–1029.
- Geiger, R.S. Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 798.
- Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (2018), 1–33.
- Kang, T. et al. Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech? *arXiv e-prints* (2022); arXiv: 2204.03262
- Korea Press Foundation. *Media users in Korea*, Korea Press Foundation (2021).
- Lee, S.H. and Kim, H.W. Why people post benevolent and malicious comments online. *Commun. ACM* 58, 11 (Nov. 2015), 74–79.

- Leshed, G. Silencing the clatter: Removing anonymity from a corporate online community. *Online Deliberation: Design, Research, and Practice* (2009), 243–251.
- Li, N. and Wu, D.D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems* 48, 2 (2010), 354–368.
- Millen, D.R. and Patterson, J.F. Identity disclosure and the creation of social capital. In *Proceedings of CHI '03 Extended Abstracts on Human Factors in Computing Systems* (2003), 720–721.
- Naab, T.K., Heinbach, D., Ziegele, M., and Grasberger, M.T. Comments and credibility: How critical user comments decrease perceived news article credibility. *Journalism Studies* 21, 6 (2020), 783–801.
- Naab, T.K., Kalch, A., and Meitz, T.G. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.
- Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.
- Pang, B. and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058* (2004).
- Rosenberry, J. Users support online anonymity despite increasing negativity. *Newspaper Research J.* 32, 2 (2011), 6–19.
- Rösner, L. and Krämer, N.C. Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media+ Society* 2, 3 (2016), 2056305116684220.
- Rowe, I. Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society* 18, 2 (2015), 121–138.
- Santana, A.D. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice* 8, 1 (2014), 18–33.
- Shin, D. Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in Korea and China. *Javnost-The Public* 26, 3 (2019), 274–290.
- Stone, A. and Potton, A. Emotional responses to disfigured faces: The influences of perceived anonymity, empathy, and disgust sensitivity. *Basic and Applied Social Psychology* 36, 6 (2014), 520–532.
- Tsang, P.P., Au, M.H., Kapadia, A., and Smith, S.W. Blacklistable anonymous credentials: blocking misbehaving users without TTPs. In *Proceedings of the 14th ACM Conf. on Computer and Communications Security* (2007), 72–81.
- Wu, T.Y. and Atkin, D.J. To comment or not to comment: Examining the influences of anonymity and social support on one's willingness to express in online news discussions. *New Media & Society* 20, 12 (2018), 4512–4532.
- Xu, G., Aguilera, L., and Guan, Y. Accountable anonymity: A proxy re-encryption based anonymous communication system. In *Proceedings of the 2012 IEEE 18th Intern. Conf. on Parallel and Distributed Systems*, 109–116.
- Zhuo, J. Where anonymity breeds contempt. *The New York Times* (2010).
- Zimbardo, P.G. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Proceedings of the Nebraska Symp. on Motivation*, University of Nebraska Press (1969).

**Minhyung Lee** is chief strategy officer at Impact AI Co., Ltd., Seoul, South Korea.

**Jinyoung Min** is an associate professor in the Department of Industrial Security at Chung-Ang University, Seoul, South Korea.

**Junyeong Lee** is an associate professor in the Department of Management Information Systems at Chungbuk National University, Cheongju, South Korea.

**Chanhee Kwak** is an assistant professor in the Department of AI Convergence Engineering at Kangnam University, Yongin, South Korea.

**HanByeol Stella Choi** is an assistant professor in the Department of Management Informationat Myongji University, Seoul, South Korea.

© 2024 Copyright held by the owner/author(s).