Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Closing Pandora's Box on Naver: Toward Ending Cyber Harassment

Nam Gu Kang*, **Tina Kuo, Jens Grossklags**

Professorship of Cyber Trust, Department of Informatics, Technical University of Munich
{namgu.kang, tina.kuo, jens.grossklags}@tum.de

Abstract

In reaction to the high-profile celebrity suicides of Sulli and Goo Hara in 2019, attributed to cyberbullying, Korea's most popular internet platform Naver introduced a range of self-regulatory measures to curtail targeted hate comments online in 2020. These regulations include removing the entertainment comment section and pseudonymizing users by revealing comment history but maintaining user anonymity. To take a closer look at celebrity cyberbullying and analyze the effects of Naver's novel self-regulatory measures, we collected data about comments and emoticons from Naver News. In the celebrity space, we find that Sulli and Goo Hara indeed received negative comments and expressions using emoticons. However, our analyses of the self-regulatory measures demonstrate that while user interaction on Naver has decreased, the percentage of negative comments has also fallen, showing that Naver's regulations have had favorable effects.

Introduction and Background

Studying South Korea's hyper-developed internet culture offers a unique opportunity to see how the future of the internet may evolve. South Korea is a fully online society: nearly 100% of the population has internet access at home and on-the-go in the form of mobile internet.[1] Supporting this densely connected society are Korea's internet portals, which offer a variety of internet services and serve as hotspots for Korean internet users. There are two main portals, Naver and Daum. This paper focuses on the effects of platform-wide self-regulatory changes between 2019 and 2020 on Naver, the most popular portal.

What makes Naver unique in the context of Korea's internet society is that the vast majority of Koreans depend on its online services. The platform is the most visited website in South Korea with a reach rate of 81.5%. The Naver app has a reach of 85.4%.[2] Naver offers many crucial services, such as online search, electronic mail, and is also the source of Naver News, the most popular news platform in South Korea by a wide margin (>30%) with a reach of 62%.[3] Naver News aggregates articles from different Korean news sources onto a single platform. Users can leave comments and emoticons directly onto news articles, making it easy for readers to contribute their opinions and to be exposed to others' thoughts.

Korean society's heavy online dependence on the internet for social connectivity is no surprise. Korea is well-known for being a collectivist society whose members value group ideals over their own and are more willing to adopt collectively beneficial values.[4] We can see the positive effects of Korea's collectivist culture in the response to the infamous Park Gyun Hye scandal of 2016, where citizens united to oust corrupt politicians in what is called the Candlelight Revolution (Chang 2020).

South Korean celebrities have massive followings partly rooted in this collectivism, which is further strengthened and enabled by platforms like Naver by providing a digital space where generations of fans can congregate online and where the praise given to these stars reaches a much higher level. For example, the K-pop group BTS not only enjoys enormous popularity, but also generates significant economic value for South Korea.[5] The other side of the coin of this heightened celebrity influence and praise is an extremely dark reality of online harassment and cyberbullying that has driven several high-profile celebrities to commit suicide and has affected millions of Naver users.

Online Harassment and Celebrity Suicide

The terminology used to describe online harassment differs across studies, which also often pursue contradictory goals (Waseem et al. 2017). On a high level, online harassment is characterized by repeated, targeted attacks on an individual (Wolak, Mitchell, and Finkelhor 2007). Different types of online harassment have been identified, including cyberbullying, hate speech, flaming, doxing, impersonating, dogpiling, and public shaming (Blackwell et al. 2017). However, the analysis of the harassment ecosystem by Pater et al.

*Nam Gu Kang and Tina Kuo share co-first authorship.

[1]https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99870&bcIdx=21930&parentSeq=21930

[2]Figures from Feb 2022: http://www.koreanclick.com/insights/service_rank.html

[3]https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

[4]South Korea scores very low on the Hofstede individualism metric and is therefore considered a collectivist society. See: https://www.hofstede-insights.com/country/south-korea/

[5]https://world.kbs.co.kr/service/contents_view.htm?lang=e&menu_cate=business&id=&board_seq=390930

(2016) illustrates that there is a lack of consistency in the definition of online harassment using a sample of 15 social media platforms and their policies.

Contributing to this definitional ambiguity, many media sources use the term "hate speech" when reporting on any content intended to produce online harassment, including South Korea's celebrity cyberbullying (Lee and Lee 2017).[6] Likewise, Waseem et al. (2017) defined hate speech as an umbrella term for expressions containing discriminatory remarks that target a generalized group or specific entity on account of race, color, national origin, sex, disability, religion, or sexual orientation. Waseem et al. (2017) further discussed the question whether an action constitutes explicit or implicit abuse, which often uses linguistically complex methods, such as sarcasm or humor. In our study of the South Korean internet culture, we consider online harassment and cyberbullying. We also use the term hate comments, when suitable.

The Naver News platform has been a hotspot for online harassment targeting celebrities, with an emphasis on cyberbullying and dogpiling. The latter term refers to collective targeting of a single individual by many users; in this case, high-profile South Korean celebrities are the targets (Blackwell et al. 2017). Korean citizens could think that celebrities are much better off than them and can be attacked online without any harm, as they do not expect their words to affect the celebrities directly (Lee and Seo 2019).[7] Additionally, online anonymity and minimal feedback on the harm inflicted by cyberbullying increases the bullies' psychological distance from their victims, dehumanizing them even further (Penny 2014).

Burdened by this unified and amplified online aggression, many celebrities have committed suicide, sparking waves of copycat suicides among private citizens (Kim and Lee 2020). One such wave of high-profile celebrity suicides (marked as P1 in Table 1) – driven in part by cyberbullying – took place in 2007-2008 and included the suicide of celebrity Choi Jin Sil, called The Nation's Actress, followed by more than a *thousand additional suicides* connected to hers, including higher rates of the use of her method of suicide.[8] To curtail such copycat suicides, reporters were asked not to report the method of suicide, and social media users can be fined $2,000 for urging someone to kill themselves.[9]

In late 2019, K-pop star Sulli committed suicide, sparking a similar copycat wave.[10] According to Kim and Lee (2020), Sulli's suicide could also be attributed to hate comments. The K-pop star Goo Hara, one of Sulli's best friends, committed suicide weeks after Sulli's death; actor Cha In Ah's suicide followed at the end of 2019 (Jung 2019). Following these celebrity suicides, the public's interest in suicides increased as can be seen in historical Google trends data.[11] In the Korean context, Kim and Lee (2020) showed that as people are exposed to celebrity suicide, they view it as a solution to their own personal struggles and worth imitating.

It must be noted that South Korea has an extremely high suicide rate relative to other OECD countries, with 24.6 deaths by suicide per 100,000 citizens (OECD 2021), representing the twelfth highest suicide rate in the OECD, according to statistics from the WHO in 2019[12]. Additionally, the high social inequality and strong emphasis on materialistic values has contributed to the extreme emotional and mental stress experienced by South Koreans (Lee and Seo 2019). Cultural norms discourage seeking help, as therapy is taboo in Korean culture, and celebrities seeking treatment risk being further criticized (Suh 2005). Instead of seeking therapy, people with lack of empathy, low self-esteem and loss of control go online to lash out at others through hate comments, perpetuating a psychological cycle of harm online (Lee and Seo 2019). Additionally, K-pop celebrities are often exploited and treated as money-making commodities subject to abuse and sexual assault by their talent agencies during their journey toward stardom.[13]

This cycle of harassment is not new for South Korea. From the infamous Dog Poop Girl[14] to the first major celebrity suicides in 2007, online harassment has repeatedly ruined lives, so much so that the government has already attempted and failed to regulate online abuse and harassment. The government created and later repealed the Network Act, which made it mandatory for all internet users to verify themselves and use their real names before commenting online (Kim 2016). Article 307 of Korea's Criminal Act has laws against defamation according to which perpetrators are liable to imprisonment or high fines.[15] In the wake of the 2019 celebrity suicides, various Korean platforms have themselves intervened with self-regulatory measures, the most notable being changes on the Naver News platform, which are the focus of our paper.

In this paper, motivated by the connection between celebrity cyberbullying and suicide in South Korea, we study whether self-regulation by platforms has a positive effect on curtailing online harassment. In a country where the internet penetration rate is as high as it is in South Korea, combined with its significant celebrity culture, the visibility and impact of celebrity cyberbullying can be extremely notable and influential. Shedding light on South Korea's understudied internet culture and investigating online harassment could pro-

[6] See, for example, https://www.thedrum.com/news/2020/09/23/facebook-youtube-and-twitter-advance-hate-speech-talks-with-brands

[7] https://mnews.joins.com/article/3321731

[8] http://english.chosun.com/site/data/html_dir/2013/01/09/2013010901227.html

[9] https://www.theguardian.com/music/2020/jan/04/i-have-reported-on-30-korean-celebrity-suicides-the-blame-game-never-changes

[10] https://www.scmp.com/lifestyle/entertainment/article/3040711/copycat-suicides-fear-korea-after-k-pop-stars-cha-ha-sulli

[11] http://english.chosun.com/site/data/html_dir/2013/01/08/2013010801097.html

[12] http://spckorea-stat.or.kr/international02.do

[13] https://emorywheel.com/underneath-the-glamour-of-k-pop-idols-a-tale-of-abuse-and-exploitation/

[14] http://legacy.www.hani.co.kr/section-005000000/2005/06/005000000200506062140001.html

[15] Korea's Criminal Act articles 307-312 https://elaw.klri.re.kr/eng_service/lawView.do?hseq=28627&lang=ENG

vide insights into how other countries might need to address online abuse. South Korea's hyper-connectivity has caused it to lead on a path that other countries will surely follow. This study's insights will contribute to our understanding of how self-regulatory interventions can be used by platforms, and our work hopefully inspires future research on South Korea's unique internet culture. To summarize, we make the following contributions:

- We highlight the importance of studying Korea's internet culture and the link between online harassment and celebrity suicide.
- Our work offers insights into the effect of pseudonymization on user behavior on a large platform, particularly, comment frequency.
- We consider the efficacy of Naver's self-regulatory interventions on cyberbullying and its implications for the general context of platform governance.

Related Work

Identifying Online Harassment in Media

Given the widespread problem of online abuse and harassment all over the world, various types of research have been developed to explore this problem space; particularly focusing on the context of social media (e.g., Mathew et al. 2019) or news platforms (e.g., Harlow 2015). These online platforms serve large-scale audiences and offer wide-ranging opportunities for user expression. Regrettably, hateful opinions incorporating malicious intent can spread especially fast on these platforms, affecting even popular and well-funded social media platforms such as Twitter and YouTube (Mathew et al. 2019; Silva et al. 2016).

Responding to the problem, several studies investigate solution or mitigation approaches (e.g., Dinakar et al. 2012). For example, over the years, many different forms of text classifiers and language models have been used to detect and eventually prevent harassment online (e.g., Saleem et al. 2017). However, despite the various efforts to rein in online abuse[16], the problem has persisted and continues to affect people across the internet and offline.[17] Moreover, the state of society at present shows negativity being transformed into real world problems.[18] Recent developments have also led to heated public discussions around controversial last-resort solutions, such as blocking individuals and organizations that are deemed to behave maliciously on their sites.[19] As such, learning from other self-regulatory approaches is helpful for addressing this problem space in the future.

Cyberbullying in South Korean Media

Following the events after Choi Jin-sil's passing, a key attempt at halting cyberbullying came from the South Korean government, which passed the Network Act (Kim 2016) (Table 1 P2). However, South Korea could never halt the online abuse problem with the Network Act and ultimately abolished it years later (Table 1 P3), because of the additional problems it created and a reduction in public support for the law as time went by (Oh 2014; Kim 2016). The need to solve the problem of cyberbullying in South Korean society again escalated after the deaths of high-profile celebrities Sulli and Goo Hara, in a repeat of the suicide wave of 2007-2008. As such, the problems with the Network Act were seen in a new light, highlighting the pressing need for solution or mitigation approaches. As was best said by the Secretary General of the Citizen's Solidarity for Human Rights in South Korea, Oh Chang-ik: "If hate speech is not truly stopped, the vicious cycle will happen over and over again" (Lee and Lee 2017).

The interactivity offered by online platforms such as social media pages and comments sections of news websites also triggers multiplier effects through the influencing of bystanders (e.g., Van den Bulck, Claessens, and Bels 2014). Being a bystander of regular celebrity harassment changes the viewer's perceptions of this behavior themselves, and even influences them to join in (Ouvrein, De Backer, and Vandebosch 2018). The extreme popularity of Naver News makes it easy for a large part of the Korean population to become an influenced bystander. However, it should be noted that the media system in South Korea also has a history of playing a powerful role in mobilizing citizens for positive social change, such as in the impeachment of corrupt President Park Geun-hye (Seo 2021). Finally, while severe cyberbullying has often been led by male aggressors focusing on the appearance and conduct of female celebrities including Sulli, the problem of cyberbullying in South Korea cuts across genders and also affects male celebrities.[20]

Public Values and Platform Intervention

In recent years, the integration and enforcement of social norms and public values in the platform ecosystem has become an increasingly important topic as more and more everyday services move onto digital platforms as part of a global trend of digital transformation. In this context, the question of who has the responsibility of enforcing public values such as privacy and transparency has been a contested issue; partly because platforms are profit-driven entities that balance commercial interests with such public values (van Dijck 2020). In response, calls have been issued for European or North American governmental institutions to act quickly to protect public values because of the reach and proliferation of digital platforms that originate from countries with different value systems that are following their own private interests (e.g., van Dijck 2020).

Data and Methods

Timeline

We created a timeline of recent, relevant events that have formed part of South Korea's fight against online abuse and

[16]https://www.thedrum.com/news/2020/09/23/facebook-youtube-and-twitter-advance-hate-speech-talks-with-brands

[17]https://cyberbullying.org/summary-of-our-cyberbullying-research

[18]https://theconversation.com/when-politicians-use-hate-speech-political-violence-increases-146640

[19]https://www.bbc.com/news/technology-55657417

[20]https://www.theguardian.com/global/2020/mar/29/behind-k-pops-perfect-smiles-and-dance-routines-are-tales-of-sexism-and-abuse

cyberbullying. We used the codes assigned in Table 1 to refer to events or time frames. The problem of cyberbullying was reignited at the end of 2019 with the prominent suicides of K-pop stars Sulli (Table 1 S1) and Goo Hara (Table 1 S2). Following these deaths, South Korean portals such as Naver put self-regulatory measures into place (Table 1 E1).

The first major change was the removal of the comment area on news articles in the entertainment section, which took place on March 6, 2020 (Table 1 E2). This intervention immediately affected all previously published as well as new articles, and effectively removed all comments from the celebrity news section.

The second major change was revealing all Naver users' comment history[21], implemented on March 19, 2020 (Table 1 E3), which was intended to foster a sense of accountability and responsibility among users. The history was made visible on a user's mandatory profile page together with a partially obscured nickname. The profile page was made easily accessible with a link next to each posted comment. We refer to this as pseudonymization rather than de-anonymization, as users' identities remain anonymous to the public, but their comment history is tracked and publicly available through their profile pages.

The Korean general elections began on April 2, 2020 and by law, all commenters were mandated to verify themselves privately with Naver before being allowed to comment during this time (Table 1 E4). As a third major change, on April 13, 2020, Naver made this identity verification mandatory, regardless of the proximity to any general election (Table 1 E5) as a further measure to improve comment quality.[22] Our research focuses around these three regulations. However, other measures were taken later in the year, including the removal of the "mad" emoticons on entertainment news articles, effective May 14, 2020.[23]

Data Collection and Methodology

All data collected for this paper were taken from Naver, as it is the largest online platform in South Korea. On Naver News, users can leave comments as well as post emoticons on articles as seen in Figure 1 (top).[24]

Naver users had the opportunity to remain largely anonymous since no real name information was made available to the public and even nicknames were being partially obscured (see Figure 1; bottom). Naver's regulation on March 19, 2020 (Table 1 E3), for the first time offered information about the commenters by revealing their comment history.

[21]https://english.hani.co.kr/arti/english_edition/e_business/933314.html

[22]It must be noted that 96% of users were already verified before this announcement https://blog.naver.com/naver_diary/221905897131

[23]https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=025&aid=0003000998&sid1=001

[24]Examples from https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=469&aid=0000501793&sid1=001

[25]Translation: Like; Warm Feeling; Sad; Mad; Want to know more (top). He's a proud father. I'm sure his wife will be a proud mother. Be happy. (bottom)

Code	Date	Event Details
P1	2007-2008	First wave of major celebrity suicides due to cyberbullying
P2	Jan 29, 2009	The Network Act stipulates that users verify themselves using their real name.
P3	Aug 23, 2012	The Network Act is abolished.
S1	Oct 14, 2019	Sulli's suicide
S2	Nov 24, 2019	Goo Hara's suicide
E1	Feb 19, 2020	Naver announces self-regulatory measures.
E2	Mar 6, 2020	Naver removes the entertainment comment section, essentially eliminating all existing comments and preventing new ones on entertainment articles.
E3	Mar 19, 2020	Naver's first action that extended across the entire platform was revealing the comment history of all users while still maintaining pseudonymous usernames. (see Figure 1 bottom)
E4	Apr 2, 2020	The Korean general elections starts. By law, all commenters must verify themselves before commenting.
E5	Apr 13, 2020	Naver announces that all commenters must verify themselves to comment regardless of proximity to any elections.

Table 1: Timeline of Relevant Events and Timeframes (**P**ast, **S**uicides, **E**vents on Naver)

Celebrity Comment Data Collection (DC1)			
Celebrity	**Comments**	**Emoticons**	**Articles**
Sulli	70258	181590	2549
Goo Hara	40264	108690	1354
IU	68390	370783	3421
Twice	105343	617395	6718
Samsung	212342	425439	6963
Individual Comment History Data Collection (DC2)			
User Discovery	**Comments**	**Commenters**	
Category	141057	300	
Corona	62566	200	
Celeb	81251	200	
Naver Article Emoticon Data Collection (DC3)			
Topic	**Emoticons**	**Articles**	
Samsung	265305	1830	
Kakao	103058	1754	

Table 2: Data Collection Overview

Three different data collections have been performed on Naver as seen in Table 2. DC1 collected comments written on celebrity articles, DC2 is a collection of the commenter's comment history, allowing insights into the effects of the March 19th regulations. DC3 included emoticons from Naver articles as a comparison set for DC2.

For DC1, we collected comments written about four celebrities as well as Samsung as a comparison set for the

Figure 1: Example emoticons available for use on Naver (top). Example of a comment on Naver (bottom).[25]

date range of January 1 to December 31, 2019. The actual data collection for the celebrities (Sulli, Goo Hara, IU and Twice) was done from February 28 to March 5, 2020, and for the comments about Samsung from March 14 to March 17, 2020. We could not retrieve a wider range of data on celebrities because of the March 6 cutoff (Table 1 E2), after which no comments on celebrities were available, as well as our technical constraints. The emoticons on entertainment articles were not directly affected by any regulation until May 14, 2020. Because we initially did not collect these, we instead gathered them from March 26 to April 2, 2020. To collect these comments and emoticons, we made use of the program `naver_news_search_scraper` posted to GitHub[26] and scripts we wrote ourselves using the Python packages `Selenium` and `BeautifulSoup`.

For DC2, we collected individual Naver users' comment histories to learn whether and how behavior had changed because of the regulation that targeted comment histories (Table 1 E3). We engaged in three rounds of scraping users' comment histories. The collected data covered the date range from January 1 to May 31, 2020.

Note that Naver has a news ranking page where the top trending news articles for the day in six different categories are listed. For the 'category' sample, we identified five trending articles posted on May 31 from each of the six different news topics. We then obtained the comment history for the aforementioned data range from each of the top 10 users for each article. This collection was performed from May 31 to June 3, 2020. For the keywords 'corona' and 'celeb', we randomly selected trending articles posted in 2019 and 2020. We chose 'corona' to find commenters who might show a tendency to comment on current issues, in this case the coronavirus, and 'celeb' was selected to identify commenters who may show an interest in celebrities, in this case particularly for Sulli, Goo Hara and Cha In Ah[27]. Comments by these users were found on articles mentioning these topics, but their histories did not necessarily show interest in this topic alone. The data collection was performed from October 17 to October 24, 2020 for 'corona' and November 14 to November 15, 2020 for 'celeb'.

For DC3, we collected data on the use of emoticons from the top news article search results for the topics of Samsung and Kakao, a company offering the popular messaging app KakaoTalk, from January 1 to May 31, 2020. More specifically, we collected the number of emoticons for five articles per day and each topic. We recorded the data for each type of emoticon. We chose emoticons, since we wanted to collect data that was not directly targeted by Naver's regulations. This gave us a dataset for comparison with the comments to gain insights into the regulations from March 19, 2020. We conducted this data collection on November 20, 2020.

BigWaveAI is a Korean AI company whose `Know Comment API` specializes in the analysis of Korean hate comments. Their team ranked first in the hate comment detection area in the AI challenge hosted by the Korean Ministry of Science and ICT and National IT Industry Promotion Agency in July 2020.[28] BigWaveAI agreed to label our comment datasets using the `Know Comment API`. We use this labeled data throughout our results section. The patented `Know Comment API` was built from a dataset of 1 million comments scraped from Naver News comments in 2019 to 2020. Then, five different workers manually labeled 500,000 of these comments as hate comments or not, with hate comments being categorized either as 'abuse' or 'offensive comments' and, discriminated by gender, age, disability, political orientation, religion, race, or appearance (Lee and Lee 2020). Among the five workers, the comments were labeled with 91% consistency. The F1-Score of the model was 84.863%.[29] Their model is additionally certified by KOIST[30] according to the Korea Laboratory Accreditation Scheme.[31]

Analysis

In this section, the various platform-wide self-regulatory steps implemented by Naver will be analyzed through changes in the collected comment and emoticon data.

Naver Entertainment Comment Section Removal

As a direct response to the suicides of Sulli and Goo Hara, Naver removed the ability to comment on news articles from their entertainment section on March 6, 2020 (Table 1 E2) and deleted all existing comments. Subsequently, only emoticons could be left as feedback on articles. On May 16, the list of emoticons was further revised to exclude the 'mad' emoticon.[32]

Analysis of Celebrity News Article Comments Our first objective was to obtain insight into the prevalence of harassing comments on topics related to Sulli and Goo Hara.

[26] https://github.com/lovit/naver_news_search_scraper

[27] Another celebrity who committed suicide shortly after Sulli and Goo Hara

[28] https://aihub.or.kr/problem_contest/5063,http://aifactory.space/aichallenge/total/search

[29] https://bigwaveai.com/, https://bigwaveai.tistory.com/

[30] Korea Information Security Technology

[31] Verified by the Korean Ministry of SMEs and Startups www.g4b.go.kr

[32] https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=025&aid=0003000998&sid1=001

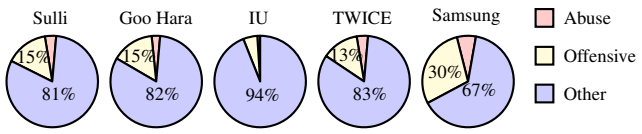

Figure 2: BigWaveAI analysis: Percentage of comments by type in 2019

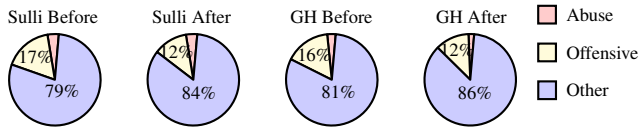

Figure 3: Average percentages of comment before and after Sulli and Goo Hara's suicides

In DC1, we collected comments and emoticons left on articles on Naver News in 2019 on the two celebrities and compared them to two female K-pop celebrities: IU, a well-known solo artist and an acquaintance of both Sulli and Goo Hara, and Twice, a high-profile girl group. We also collected comments left on articles concerning Samsung Electronics as a comparison set. These comments were then analyzed with the BigWaveAI `Know Comment API` and labeled 'abuse', 'offensive' or other.

Inspecting the values in Figure 2, we can observe that the majority of comments posted in relation to all artists were not categorized as abusive or offensive comments. However, comments relating to Sulli, Goo Hara and Twice were labeled as offensive at rates of 15%, 15%, and 13%, respectively. While these percentage figures constitute only half of the percentage of offensive comments targeting Samsung, it remains a considerable amount of received harassing comments when directed at human beings.

The relative prevalence of offensive comments directed at Sulli and Goo Hara shown in Figure 3 suggests that the percentage of offensive comments was somewhat higher before their deaths.

The BigWaveAI analysis also included additional data on individual comments *if* the comments were labeled as abuse or offensive. These data were related to discrimination based on gender, age, disability, political orientation, religion, race, and appearance. Using these categorizations, we see a difference between the solo artists relative to Twice. The most common type for Sulli, Goo Hara and IU was gender-related at 15.9%, 16.6%, and 10.8%, respectively. In contrast, while Twice also received 15.9% gender-related comments, the highest percentage was race-related comments at 23.8%. Samsung's highest percentages were political and related to race at 37.9% and 16.3%, respectively.

Examining the emoticon feedback that the artists received in Figures 4 and 5, we see that Sulli and Goo Hara both received more 'mad' feedback than any other category in 2019 before their suicides. After their deaths, the 'sad' emotions became the majority. This differs from the other two artists, where 'like' or 'warm' emoticons, indicating that

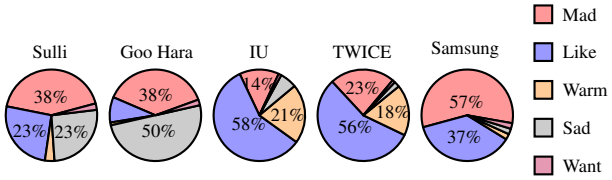

Figure 4: Percentages of emoticon use in 2019

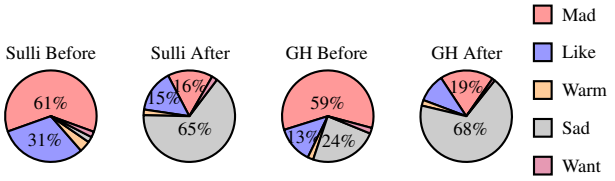

Figure 5: Percentages of emoticon use before and after Sulli and Goo Hara's suicides

commenters were touched by the article, were consistently more frequent than all other emoticon types.

> Finding 1: Sulli and Goo Hara received substantial expressions of negative sentiment on Naver News with a notable percentage of abuse and offensive comments and with the largest percentage of 'mad' emoticons left on their articles before their deaths.

Mandatory Public Comment History

In this section, we examine the March 19 (Table 1 E3) intervention making comment histories mandatory and publicly available, however, in a pseudonymous fashion.[33]

Comparison to Data from Naver Datalab In DC2, we collected 700 different users' comment histories to understand the impact of the March 19 regulation changes. These commenters came from the 'category', 'celeb', and 'corona' datasets. To get an initial understanding of the data, we compared the number of comments to data from Naver Datalab[34], a website where Naver also communicates how many comments were posted to the news platform per day. Our daily data, on average, represented 0.42% of the total daily comments written on Naver.

Decrease in Comment Volume Post-Pseudonymization First, we looked at the effects of the regulation to the total amount of comments posted on Naver to determine whether the intervention would be associated with users commenting less. Causal inference evaluates observational data pre- and post-event for a significant causal effect from a particular event. The traditional differences-in-differences causal inference analysis compares differences between prior and

[33] https://blog.naver.com/naver_diary/221815294149, https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=003&aid=0009762720&sid1=001

[34] https://datalab.naver.com/commentStat/news.naver

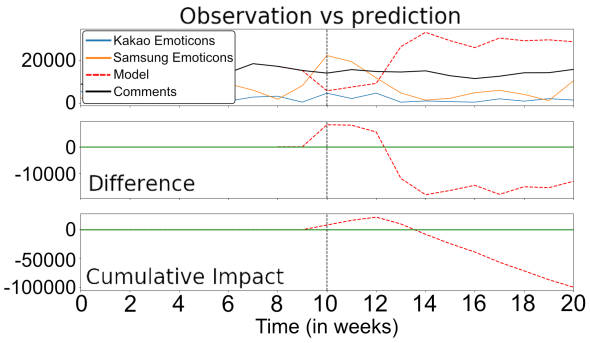

Figure 6: Causal impact of total amount of weekly comments with emoticons

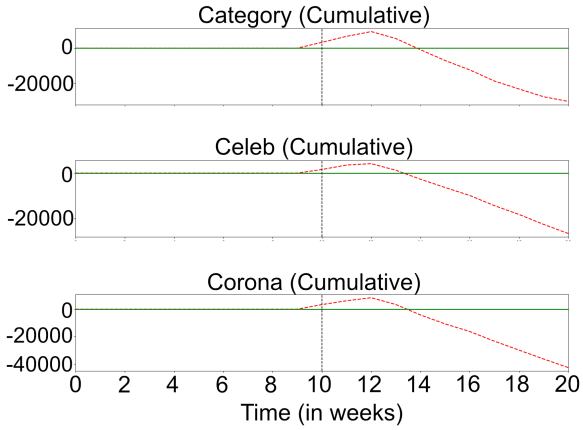

Figure 7: Causal impact of total amount of weekly comments from the three datasets with emoticons

posterior data between groups that do and do not experience the intervention (Bertrand, Duflo, and Mullainathan 2004). The causal analysis method uses the entire dataset and treats each data point separately (Brodersen et al. 2015). Following the latter method, we use the Python package `causal_impact` uploaded on GitHub for the analysis and creation of the causal impact graphs.[35]

We investigated whether the amount of comments on Naver were affected by the regulation to show commenters' histories (Table 1 E3). To apply the methodology, we assumed that the other method for users to express their emotions and opinions (i.e., emoticons) would not be affected as these would still not be shown on the users' histories. In addition, Naver had no regulations on the emoticons in the normal news section from where we collected our emoticon dataset in DC3; for the topics Samsung and Kakao (see Table 2). In a causal impact analysis, we then need to give the date of intervention. Using this date, the causal impact analysis shows us what the affected dataset would have been like if there was no intervention and shows us the difference from this prediction to the actual observation. In our analysis, we made use of the category, celeb and corona comment datasets from Jan 5 to May 30 as our affected variable. Our unaffected variable are the emotions pulled from Samsung and Kakao articles, likewise from Jan 5 to May 30. We grouped the comment and emoticon numbers by obtaining the total number of comments and emoticons per week in relation to the fluctuations in daily data. We set week 1 as beginning on Jan 5, the first Sunday of 2020 and ending on May 30 in Week 20. Week 10, the week of March 19, was set as the intervention date. A causal impact analysis was run with an aggregated set of our category, celeb and corona comments and then individually for each partial dataset.

Using the full dataset in Figure 6, we can observe that the prediction, the dotted red line labelled "model", expected more comments than were actually posted, i.e., the black line labelled "Comments". This results in a decrease in the cumulative impact, or sum of all differences, after March 19, as illustrated in the bottom two graphs in Figure 6. We add robustness to our analysis by conducting it for the three datasets individually, which also showed a decrease in comments, post-regulation, as seen in the cumulative impact graphs in Figure 7. Overall, this is indicating that the commenters in our dataset wrote less comments after the regulation went into effect.

> Finding 2: The analysis of the collected comments on Naver showed a decrease in the number of comments after the regulation to reveal comment histories went into effect.

Decrease in Negativity in Comments Post-Intervention To examine the impact of the regulations on the comments themselves, we split our datasets into before and after the regulation took effect and observed the percentage of 'abuse' and 'offensive' comments relative to the total dataset. Our dataset is centered around March 19. Hence, we first split the data into January 5 to March 18, 2020 and March 19 to May 31, 2020. (See Table 3 for an overview.) In Table 4, we then set a before and after window of two weeks around the following three intervention dates: the March 6 regulations (Table 1 E2)[36], the March 19 regulations (Table 1 E3), and the April 2 regulations (Table 1 E4) to understand how the comments changed as a result of those interventions.

We conducted paired sample t-tests[37] to check for statistical significance regarding the before and after data using the number of comments labeled as abuse or offensive. For the data in Table 3, we set our alpha level to 0.05 and the t-critical value to 1.67 given the data from 73 days before and after March 19. The statistical significant pairs are underlined in Table 3.[38] For the data in Table 4, our alpha level was set again to 0.05 and t-critical value to 1.77 given the

[35] https://github.com/tcassou/causal_impact

[36] While the dataset does not contain any entertainment section comments, we can still see the change in user behavior over other sections on Naver.

[37] https://dfrieds.com/math/dependent-samples-t-test.html

[38] One exception was found for the abuse comments before and after in the 'category' dataset.

DC2		Abuse		Offensive	
Dataset	**Type**	**Before**	**After**	**Before**	**After**
Category	**Abs.**	4216	4397	21160	24389
	Percent	7.24%	5.54%	20.96%	20.11%
Celeb	**Abs.**	2178	1559	11758	8714
	Percent	6.79%	5.46%	36.35%	30.71%
Corona	**Abs.**	2684	1779	15478	11247
	Percent	6.51%	4.72%	37.56%	29.85%

Table 3: BigWaveAI Commenter Datasets Before and After March 19

DC2		Abuse		Offensive	
Dataset	**Type**	**Before**	**After**	**Before**	**After**
Category	**Mar 6**	8.14%	6.84%	38.42%	28.38%
	Mar 19	7.23%	5.64%	34.43%	31.05%
	Apr 2	5.63%	4.50%	29.83%	33.38%
Celeb	**Mar 6**	7.38%	5.35%	40.79%	24.87%
	Mar 19	5.72%	4.46%	33.83%	28.80%
	Apr 2	4.26%	4.89%	28.92%	30.14%
Corona	**Mar 6**	7.01%	4.41%	38.97%	26.30%
	Mar 19	4.90%	3.73%	35.03%	26.83%
	Apr 2	3.59%	3.60%	28.52%	28.70%

Table 4: BigWaveAI Commenter Datasets two Weeks Before and After Events

14 days before and after the event dates. Again, we checked for statistical significance of data using the number of comments labeled as abuse or offensive 14 days before and after a certain date. The dates of the regulations that led to a statistically significant difference in comment history are underlined in Table 4.

The statistical test results as well as the overall percentage values in Table 3 vividly illustrate the decrease in both abuse and offensive comments across the datasets. For the before and after (two-week window) data in Table 4, we see that the March 6th and 19th regulations were more impactful in causing changes to commenters' behaviors. In contrast, we can observe that the private identification verification had less of an impact.

Naver's statement that 96% of all users were already verified[39] also helps us understand these results. Given this fact, the identification verification regulation should not have affected the platform the way the other regulations did, which had a platform-transforming character.

> Finding 3: We observe a decline of abuse and offensive comments after Naver's major regulations to remove the entertainment comment section and the revelation of user comment histories.

Increase in Positivity in Overall Comments Sentiment To solidify the findings from the previous section, we performed another causal impact analyses on the commenter datasets (similar to the analysis done for Finding 2). There,

[39]https://blog.naver.com/naver_diary/221905897131

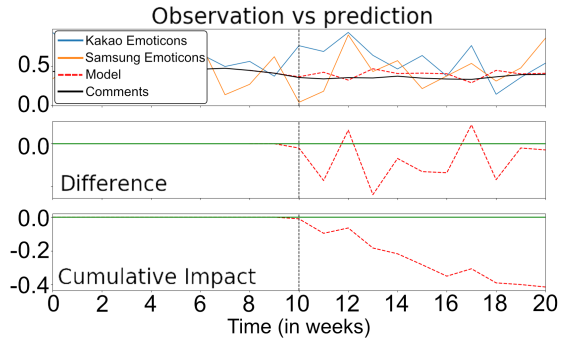

Figure 8: Causal impact of total weekly abuse/offensive comments with the 'mad' emoticon

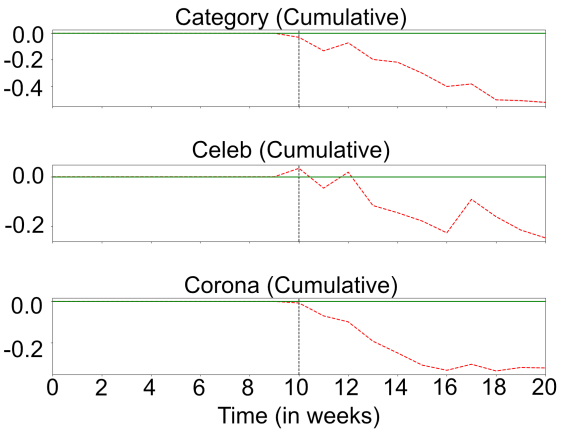

Figure 9: Causal impact of weekly abuse/offensive Comments from the three datasets with the 'mad' emoticon

we found that there were less comments posted overall after Naver's regulation to reveal comment histories.

In this section, we aim to understand if the percentage of comments labeled as abuse or offensive also decreased. For this, we changed the affected variable from the total amount of comments to the percentage of comments labeled as abuse or offensive compared to the total dataset. To match the change to the comments, we also changed the non-affected emotions to be the percentage of the 'mad' emoticon compared to the rest of the emotions in the Samsung and Kakao datasets. Here, we reused the data groupings and took the total number of 'abuse' and 'offensive' comments/'mad' emotions per week divided by the total number of comments/emoticons per week. Again as in the previous causal impact analysis, we took week 10, the week of March 19th, as the event date.

When examining the results for the total dataset in Figure 8, we learn that while there are weeks where we had a higher percentage of abuse and offensive comments compared to prediction, overall in the cumulative impact the percentage of abuse and offensive comments is reduced. Similar results are observable for the three individual comment history datasets in Figure 9.

Finding 4: The regulation that revealed comment histories caused a decrease in the percentage of negative comments and a corresponding increase of the percentage of positive ones.

To summarize, with Finding 1 we confirmed that Sulli and Goo Hara faced substantial negativity on Naver, which likely motivated the self-regulatory efforts on Naver. Finding 2 demonstrates that comment volume decreased after Naver's regulation to reveal commenter history. In addition, Findings 3 and 4 illustrate that the percentage of comments labeled as abusive or offensive also decreased suggesting desirable effects of the interventions, at least in the short run.

Limitations

In this paper, we acknowledge several limitations. As mentioned previously in the data collection section, we were only able to compare comments on Sulli and Goo Hara with two other celebrities. This was because we only began scraping comments after the announcement of the removal of the ability to comment on the entertainment comment section and, hence, we were unable to take the data for more celebrities because of time constraints. Additionally, we only picked two other Korean celebrities who seemed to form an adequate comparison for Sulli and Goo Hara. As such, the analysis of the celebrity comments would certainly gain robustness if data for a wider variety of celebrities could be incorporated, but this data is not accessible anymore.

When further examining the effects of the regulations, we derived findings from a sample of commenters on Naver. We consider that this sample from Naver was sufficient to obtain adequate insights into the effects of the regulation toward commenters; however, a longer timeframe and larger sample may be beneficial. Further, individual commenter behavior analysis could produce more in-depth results, which however goes beyond the scope of this paper. Also, for the causal impact analysis findings, we had to make assumptions about which variables were and were not affected by interventions. To provide additional evidence, we complemented our analysis with basic statistical tests. Finally, we did not identify other factors outside of Naver's regulations for the platform. Considering any further influences in the analysis would be beneficial for further research.

Despite the limitations of our work, we argue that our paper provides important first insights into Naver's platform suggesting some notable effects of Naver's attempts to stop online harassment. We hope that this will prompt further research into the South Korean media landscape and its attempts to combat online abuse and cyberbullying, and to complement more frequent research on Western social network sites.

Discussion

Effects of Pseudonymization on User Behavior

One commonly discussed topic in internet research is the impact of de-anonymization on user behavior, as the feeling of anonymity could be a driving factor for users' aggressive expression online because actions cannot be attributed directly to them (Wallace 2015). In an attempt to stop hate comments in 2007 in the wake of major celebrity suicides, Korea's government promulgated the Network Act (Table 1 P2), which made it mandatory for all Korean users to verify themselves before being able to comment online (Oh 2014). However, this first step towards de-anonymization was heavily criticized because of the problems it created and was later repealed (Kim 2016; Park and Greenleaf 2012). For example, Leitner criticized the Network Act saying that it limited expression online (Leitner 2011, 2009). Kim (2016), in his study about the Network Law, explained how the way the system handled real name information was flawed and that it did not impede hate comments (Park and Greenleaf 2012). Additionally, some users apparently just moved over to foreign platforms (Caragliano 2013).

Naver's pseudonymization regulations show a different approach to online de-anonymization, which learned from the shortcomings of the Network Act. Personal information is not exposed to the public and users are not de-anonymized. Instead, a sense of reputation and responsibility for the user behind the account is created by revealing the account comment history. Users remain anonymous, but their comment histories are made visible to the public under their partly obscured usernames, or online pseudonyms. This approach is related to insights about the online disinhibition effect, which highlights that anonymity causes people to act online as they would not in the physical world (Suler 2004). For example, a study on Dutch news sites showed that after de-anonymizing comments by linking real-name Facebook profiles, the quality of comments rose, although fewer people commented overall (Hille and Bakker 2014). We see this reflected in Finding 2, namely, that although overall fewer comments were being posted, there was a reduction in comments labeled as abusive or offensive, as seen in Findings 3 and 4. Complementary to our results, the Korean Press Foundation found that there was an increase in the average number of characters in the posted comments. The organization concluded that the comment section was likely improving in quality.[40]

Before Naver's self-regulation, it was impossible to know whether a particular commenter often posted malicious comments. The anonymous comment culture allowed users to write virtually anything that they wished on Naver and to essentially begin fresh with every comment because no one could follow any given user's tracks online. Further, as every posting began a commenter's interaction anew, all commenters were placed on an equal playing field, and each comment presented was likely valued equally. After the comment histories were revealed, news sources highlighted cases were commenters had been essentially catfishing online, for example, pretending to have different identities when posting comments.[41] After mandatory comment his-

[40] https://www.kpf.or.kr/commonfile/fileidDownLoad.do?file_id=00050260E95BA256F2F1C8A426A451ED&board_id=246&contents_id=60ba51802ff04bbc872275a785137b69

[41] https://www.yna.co.kr/view/AKR20200327192300017,

tories were enabled, these users were essentially forced to stop their negative practices online and stick with one version of their story. Koreans place great emphasis on not losing face and on maintaining a good reputation, which could be used to halt hate comments (McDonald 2011).[42] Mandatory comment histories allow users to obtain a better idea of the context in which a negative or even hateful comment is made and provide an environment for more reliable content by allowing users to ignore consistently malicious users or trolls (Wu and Atkin 2017).[43]

Control of User Expression on Naver

In addition to the pseudonymization of all Naver users, Naver significantly changed the way that their entertainment section functioned by removing the comments section (Table 1 E2) and changing the emoticons to exclude the mad emoticon as a direct response to the celebrity deaths of 2019. While pseudonymization still allows users to decide how they want to use Naver News as a platform, the changes in the entertainment section set fixed rules and limits on what users could do and share.

The first regulation prevented comments from being written, which can be interpreted as a suppression of user's ability to publicly express themselves and discuss on the platforms. However, without regulation there is the risk that comments, thoughts and ideas from cyberbullies could be transferred from an abusive commenter to other readers (Ferráns, Selman, and Feigenberg 2012; Lee, Jang, and Chung 2020). Naver's wide reach entails that hate comments have the potential to quickly reach a majority of the Korean population. Additionally, Korea's collectivist society leads to further problems as individuals' ideals could quickly reflect those of the group (Ahn 2011) and reach potentially extreme levels (Kim 1993). When comments are able to create negative ideas about celebrities in the minds of others and make them into the group's "ideas", they are no longer just cyberbullies' thoughts but may reflect a larger part of the entire Korean community. Sulli mentioned shortly before her suicide that she felt watched and attacked by everyone, who believed everything they read about her online.[44] Other celebrities have made similar observations.

The removal of comments on Naver eliminated an avenue for cyberbullying from Naver's entertainment section, but not from Korean society as a whole. Under the Network Act, one of the points of failure was that some users simply moved to alternative platforms (Kim 2016; Caragliano 2013). News reports emerged of cyberbullies simply joining different social media platforms such as Instagram.[45] On the one hand, some of these alternative platforms do not have

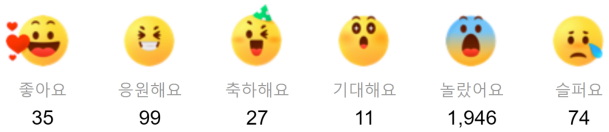

Figure 10: The new emoticon system on the entertainment section excluding the 'mad' emoticon [46]

nearly the same audience and spread potential as Naver in South Korea. On the other hand, we see that hateful comments will continue even if major platforms regulate online harassment. Even worse, if such comments are included as 'shocking examples' or gossip in news articles, their reach can even increase.

Nevertheless, it is illustrative to observe that a level of control can be imposed on how information is delivered. For example, in Finding 3, we suggested a notable decrease in the percentage of offensive comments following the March 6th regulation to remove the entertainment comment section. In a paper comparing Jonghyun's and Sulli's suicides, Kim and Lee (2020) pointed out that celebrity suicides have different effects on society depending on how media reported on the suicides: for example, focusing more on hate comments that caused suicides then decreases copycat suicides in comparison to focusing on the suicide itself. Seeing that the topic of suicide and the prevention of further cases can be influenced by the media, we can hope that the spread of hate comments can also be controlled reasonably.

The second change was the change in emoticons that could be posted on news articles as seen in Figure 10.[47] The previous emoticon system on the entertainment section seen in Figure 1 which included the 'mad' emoticon may have also fueled the further spread of hateful opinions. We saw in Finding 1 how Sulli and Goo Hara received more mad emoticons compared to the other artists. Removal of this emoticon prevented users from expressing this emotion with regard to an article, as Naver's new range of available emoticons was more neutral or positive. However, this leads to the case of Naver essentially controlling what people should think of the entertainment section articles as users cannot express a broader range of emotions and can only use a restricted range of emoticons, leading to the possibility of further problems such as critical opinions and comments being crowded out and being unaddressed.

One final consideration to keep in mind when looking at regulations on Naver is the question: who is the ultimate beneficiary of Naver's actions. Following Shin (2019), Naver has a history of manipulating how they present news or information in relation to users with specific agendas. This may motivate a more critical narrative which inquires whether Naver's regulations were truly a response to the celebrity suicides or a face-saving move to solve a problem

https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=421&aid=0004591056&sid1=001, http://news.tvchosun.com/site/data/html_dir/2020/04/05/2020040590109.html

[42]https://www.youtube.com/watch?v=HxSNqUvZtUY

[43]https://www.viki.com/tv/36999c-77-billion-in-love

[44]https://www.viki.com/tv/36655c-night-of-hate-comments

[45]https://view.asiae.co.kr/article/2020062612182682857, https://www.seoul.co.kr/news/newsView.php?id=20200831010005&wlog_tag3=naver

[46]Translation: Like ; Cheer On ; Congrats ; Looking forward to it ; Surprised ; Sad

[47]Examples from https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=469&aid=0000501793&sid1=001, https://entertain.naver.com/ranking/read?oid=076&aid=0003681458

they partly helped to create in the first place. We did not investigate this further in this paper, but we think that additional research into potential user manipulation by Naver's regulations would be valuable.

Platform Responsibility and Governance

As South Korea's social infrastructure largely exists online on platforms like Naver, these online platforms' policies and governance have an astoundingly large effect on society as a whole. However, observers have commented that companies that run mega-platforms such as Facebook and Google often try to shirk public and legal responsibility for what occurs on their platforms (van Dijck 2020). We have seen the unprecedented societal consequences that the spread of misinformation has had on society at large, such as in the 2016 presidential election in the United States (Allcott and Gentzkow 2017). Nonetheless, the regulation of problematic third-party content such as hate comments and misinformation is increasingly dependent on the effective involvement of private entities as they are in the best position to take steps for improvement on their own platforms.[48] Unlike Section 230 in the U.S. Communications Decency Act, which partly shields companies from responsibility, Article 44(3) of Korea's Information and Communications Network Act encourages companies to proactively censor problematic content posted on their websites, threatening years in prison or a fine of up to 50 million Korean Won ($9,000), even if the allegedly defamatory statements are proven to be true.[49] Our results hint at improvements made by Naver as they likely decreased the volume of hate comments by removing the comment options for the particularly heated entertainment section and by introducing user history transparency. Revealing commenter histories also meant that users were less likely to write comments that they could later disavow.

In this context, Shin (2019) emphasized not only the imperative need of algorithmic technology in the Korean media sector, but also the importance of fairness, accountability, and transparency that must accompany the implementation of such technologies. The removal of an aggressive, hate-filled comment section by Naver is in line with results of prior findings that bystanders are likely to engage in a similarly aggressive fashion online if that is seen as the discussion forum's norm (Cicchirillo, Hmielowski, and Hutchens 2015). Our case study of South Korean media shows that platform-led regulation does have a positive impact on controlling cyberbullying on their own platform.

Conclusion

Through our findings in the celebrity cyberbullying space, we provided further evidence that South Korea and Naver are suffering from a problem with aggressive and hateful comments about celebrities. However, the observed instances of celebrity suicides eventually pushed for a change. The results of our analysis of the regulations enacted by Naver show that while there are negative side effects on the platform's performance such as decreased user interaction, the frequency of abuse and offensive comments also decreased. Overall, the changes on Naver News seem promising as Naver's self-regulations did not suffer from similar pitfalls as the Network Act, giving users an incentive to maintain their reputation on the platform without revealing users' personal information. Further research and work about Naver and South Korea more generally is surely needed as new challenges arise. Our work showcases a rare example of a private platform company taking the initiative to curtail cyberbullying with a positive impact, but with a cost of reduced overall activity on the platform. We hope that this case study on Naver will serve to inform and encourage future platform interventions against online harassment.

[48] https://www.washingtonpost.com/opinions/why-big-techs-attempt-to-stifle-free-speech-could-be-futile--or-worse/2021/01/11/e912acac-5432-11eb-a931-5b162d0d033d_story.html

[49] https://lawless.tech/internet-laws-south-korea/

Acknowledgments

We thank the reviewers for their insightful comments. We gratefully acknowledge support from the Institute for Ethics in Artificial Intelligence (IEAI) at the Technical University of Munich. We are very thankful for Hee Jun Lee and the team at BigWaveAI for informing and labeling our comment dataset with their Know Comment API.

References

Ahn, D. 2011. Individualism and collectivism in a Korean population. *Scripps College Senior Theses*, Number 107.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2): 211–36.

Bertrand, M.; Duflo, E.; and Mullainathan, S. 2004. How much should we trust differences-in-differences estimation? *The Quarterly Journal of Economics*, 119(1): 249–275.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2017. Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–19.

Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9(1): 247–274.

Caragliano, D. 2013. Real names and responsible speech: The cases of South Korea, China, and Facebook. https://www.yalejournal.org/publications/real-names-and-responsible-speech-the-cases-of-south-korea-china-and-facebook. Accessed: 2022-04-02.

Chang, H.-J. 2020. South Koreans worked a democratic miracle. Can they do it again? https://www.nytimes.com/2017/09/14/opinion/south-korea-social-mobility.html. Accessed: 2022-04-02.

Cicchirillo, V.; Hmielowski, J.; and Hutchens, M. 2015. The mainstreaming of verbally aggressive online political behaviors. *Cyberpsychology, Behavior, and Social Networking*, 18(5): 253–259.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, pre-

vention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3): 1–30.

Ferráns, S. D.; Selman, R.; and Feigenberg, L. F. 2012. Rules of the culture and personal needs: Witnesses' decision-making processes to deal with situations of bullying in middle school. *Harvard Educational Review*, 82(4): 445–470.

Harlow, S. 2015. Story-chatterers stirring up hate: Racist discourse in reader comments on US newspaper websites. *Howard Journal of Communications*, 26(1): 21–42.

Hille, S.; and Bakker, P. 2014. Engaging the social news user: Comments on news sites and Facebook. *Journalism Practice*, 8(5): 563–572.

Jung, H.-M. 2019. Copycat-suicides fear in Korea after K-pop stars Cha In-ha, Sulli, Goo Hara die in the space of two months. https://www.scmp.com/lifestyle/entertainment/article/3040711/copycat-suicides-fear-korea-after-k-pop-stars-cha-ha-sulli. Accessed: 2022-04-02.

Kim, K. 2016. Korean Internet and 'real name' verification requirement. *Korea University Law Review*, 20: 87–103.

Kim, K.-O. 1993. What is behind face-saving in cross-cultural communication. *Intercultural Communication Studies*, 3(1): 39–48.

Kim, Y.; and Lee, S.-B. 2020. The semantic network analysis of celebrity suicide news: The case study of SHINee Jonghyun and F(x) Sulli. *Legislation and Policy Studies*, 12(2): 339–369.

Lee, E.-J.; Jang, Y. J.; and Chung, M. 2020. When and how user comments affect news readers' personal opinion: Perceived public opinion and perceived news position as mediators. *Digital Journalism*, 9(1): 42–63.

Lee, H.-J.; and Seo, J.-A. 2019. You could be a real hero when you stop being a keyboard warrior. http://www.hanyangian.com/news/articleView.html?idxno=959. Accessed: 2022-04-02.

Lee, S.-K.; and Lee, Y.-E. 2017. Prevalence of hate speech in the Korean society. http://www.hanyangian.com/news/articleView.html?idxno=800. Accessed: 2022-04-02.

Lee, W.; and Lee, H. 2020. Bias & hate speech detection using deep learning: Multi-channel CNN modeling with attention. *Journal of the Korea Institute of Information and Communication Engineering*, 24(12): 1595–1603.

Leitner, J. 2009. Identifying the problem: Korea's initial experience with mandatory real name verification on Internet portals. *Journal of Korean Law*, 9: 83–108.

Leitner, J. 2011. To post or not to post: Korean criminal sanctions for online expression. *Temple International and Comparative Law Journal*, 25(1): 43–77.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, 173–182.

McDonald, M. 2011. Stressed and depressed, Koreans avoid therapy. https://www.nytimes.com/2011/07/07/world/asia/07iht-psych07.html. Accessed: 2022-04-02.

OECD. 2021. Suicide rates (indicator). https://data.oecd.org/healthstat/suicide-rates.htm. Accessed: 2021-01-04.

Oh, Y.-H. 2014. *A study of Internet real name policy change*. Master's thesis, Seoul University.

Ouvrein, G.; De Backer, C. J.; and Vandebosch, H. 2018. Joining the clash or refusing to bash? Bystanders reactions to online celebrity bashing. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12(4): Article No. 5.

Park, W.-i.; and Greenleaf, G. 2012. Korea rolls back 'real name'and ID number surveillance. *Privacy Laws & Business International Report*, 119: 20–1.

Pater, J. A.; Kim, M. K.; Mynatt, E. D.; and Fiesler, C. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 369–374.

Penny, L. 2014. *Unspeakable things: Sex, lies and revolution*. Bloomsbury Publishing USA.

Saleem, H. M.; Dillon, K. P.; Benesch, S.; and Ruths, D. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Seo, S. 2021. South Korea's Watergate moment: How a media coalition brought down the Park Geun-hye government. *Journalism Practice*, 15(4): 526–543.

Shin, D. 2019. Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in Korea and China. *Javnost – The Public*, 26(3): 274–290.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, 687–690.

Suh, G.-H. 2005. Mental healthcare in South Korea. *International Psychiatry*, 2(7): 10–12.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & Behavior*, 7(3): 321–326.

Van den Bulck, H.; Claessens, N.; and Bels, A. 2014. 'By working she means tweeting': Online celebrity gossip media and audience readings of celebrity Twitter behaviour. *Celebrity Studies*, 5(4): 514–517.

van Dijck, J. 2020. Governing digital societies: Private platforms, public values. *Computer Law & Security Review*, 36: Article No. 105377.

Wallace, P. 2015. *The psychology of the Internet*. Cambridge University Press.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84.

Wolak, J.; Mitchell, K. J.; and Finkelhor, D. 2007. Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *Journal of Adolescent Health*, 41(6): S51–S58.

Wu, T.-Y.; and Atkin, D. 2017. Online news discussions: Exploring the role of user personality and motivations for posting comments on news. *Journalism & Mass Communication Quarterly*, 94(1): 61–80.