Kaneyasu, Keisuke; Koguchi, Teppei

Conference Paper

Content moderation preference analysis by digital platforms: Based on Japanese case

24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024

Provided in Cooperation with:
International Telecommunications Society (ITS)

Suggested Citation: Kaneyasu, Keisuke; Koguchi, Teppei (2024) : Content moderation preference analysis by digital platforms: Based on Japanese case, 24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/302461

Content moderation preference analysis by digital platforms: Based on the Japanese case

Keisuke Kaneyasu, Teppei Koguchi

Keywords: digital platform, content moderation, YouTube

1. INTRODUCTION

Digital platforms are receiving increasing attention. By 2022, Internet access will continue to expand, reaching 66% of the population. Digital platforms can provide services worldwide, except for a few countries. Consequently, several digital platforms with large user pools were developed. Digital platform users browse content on digital platforms. In addition, users use them as places to create and express content, such as text, images, and videos. Currently, because many users use a platform with different values, some behaviors are offensive to others. Therefore, digital platform operators consider what and how much to allow and what and how much to regulate. As a result of their deliberations, they operate with restrictions on services while establishing legal regulations, such as the terms of use created by the operators. In this case, the larger the digital platform becomes, the greater its impact as a speech space becomes, so attention will be paid to the criteria for controlling what kind of content is considered an infraction and what kind of content is hidden.

However, the definition of content moderation remains unclear. The EU's Digital Services Act defines content moderation in this manner. "Content moderation" means the activities, whether automated or not, undertaken by providers of intermediary services that are aimed, in particular, at detecting, identifying, and addressing illegal content or information incompatible with the terms and conditions provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account.

In addition, Santa Clara Principles 2.0 also defines content moderation as follows. The term "content" refers to all user-generated content, paid or unpaid, on a service, including advertising. The terms "action" and "actioned" refer to any form of enforcement action taken by a company with respect to a user's content or account due to non-compliance with their rules and policies, including (but not limited to) the removal of content, algorithmic downranking of content, and the suspension (whether temporary or permanent) of accounts. Recently, from the perspective of FactCheck, there have been ideas such as placing a checkmark on the content, which has transformed how to deal with this issue. Therefore, several interpretations have been proposed. In this study, to emphasize the user's point of view, we define it as follows. Content moderation is a measure taken by a platform to render a user's content invisible to viewers. This includes the act of deleting the target account and hiding all posted content at once, as well as the act of not accepting any further posts.

2. PREVIOUS STUDY

Along with the criticism that digital platforms have problems with antitrust laws, freedom of expression, and so on, there are opinions that they should fulfill more social responsibilities. For example, the Democratic Party in the US has expressed hope that technology companies and social media will take on greater responsibility than they do now. There have been several proposed revisions to the Communications Decency Act in the US that would significantly affect content moderation. Tachibana (2022) analyzed Section 230 of the Communications Decency Act in the US in light of former president Donald Trump's series of information deliveries and platformers' responses at that time. Through this analysis, he made the following recommendations: "Imposing obligations on providers raises competition law issues. Government involvement can also decrease privacy and freedom of expression. Therefore, transparency is recommended."

Regarding information on digital platforms, Mizutani (2022) provided an analysis of the Florida law that prohibits digital platforms from moderating content in the US. In his analysis, he

noted that the US District Court for the Northern District of Florida expressed “the truth somewhere in between” as to whether social media should be treated like newspapers and other media or as common carriers. He also argues that the content circulating on the platform is more akin to a product that is mechanically generated in a factory than to information from a news editorial office. Although content moderation has received much attention, Jialun et al. (2022) pointed out that most content moderation research evaluates issues from a certain perspective and that we should be more aware of the inherent trade-offs involved.

In Japan, there are also a lot of legal and academic approaches. Under Japanese law, the Provider Liability Limitation Act, there is room for liability for damages for a breach of the provider’s duty to remove illegal content known to the provider if the provider itself does not remove it. On the other hand, it is not clear how to deal with content that is not immediately illegal (Watanabe, Umemoto, and Imamura, 2021). As part of the government’s study on content moderation in Japan, the Ministry of Internal Affairs and Communications (MIC) released the final report of its study group on platform services in February 2020. The report provides a summary and directions for issues and concerns that have arisen as platform services have expanded their market presence. The report also states that the private sector should take voluntary measures against fake news and disinformation since legal regulations may have the effect of atrophying free expression, lack of effectiveness, and arbitrary operation.

Another study focused on a specific area is that of Yamaguchi (2015). He conducted a survey of flaming incidents and found the following. Only 1.5% of users are involved in slander and flames on the Internet. The younger the youth, the more they feel that the Internet is a good place to say what they want, and the more they feel that it is okay to blame others.

The following is the “Results of a Questionnaire Survey on the Circulation of Slanderous and Defamatory Information on the Internet” released by Mitsubishi Research Institute (2022). This report was released by the “Working Group on Countermeasures against Illegal and Harmful Information, Including Slanderous and Defamatory Information,” established by the MIC within the study group on platform services. According to this report, SNS(Social networking service) users are in the following situations: A. 65.1% of users have witnessed posts that hurt others (slander). B. Less than 20% (18.3%) of users have been victims of “posts that hurt others (slander)” in the past year. C. 35.3% of the users “wanted to use the functions but did not know how to use them” or “did not know that the functions were available” regarding the safety and security functions such as mute and block. Toriumi and Yamamoto (2022), who examined the effects on users, compared information acquisition to health and nutrition and described a state of constant immunity as “informational health.” They also recommended a well-balanced acquisition of information to counteract the information eclipse caused by the attentional economy.

Among the many studies conducted in Japan are government-led countermeasures against slander and fake news. These issues, such as political division and international security, are sometimes discussed from a public perspective. However, as noted in the MIC report, it is desirable to consider this from a user’s perspective. From the user’s perspective, we assume that some users want action against content that is not illegal, whereas others believe that it is acceptable to the extent that it is not illegal but detrimental to them. In Japan, no survey has been conducted from the perspective of what kinds of moderation users want for non-illegal content, and it is unclear what kinds of standards users want.

3. PURPOSE OF THE ANALYSIS AND CONDUCT OF THE SURVEY

Following the Japanese government's approach, users should get internet literacy as possible, but the argument that digital platforms should perform content moderation is leading. However, there are problems with this approach. When digital platforms overdo content moderation, users may lose their right to know and their freedom of expression. Certainly, if users are not exposed to Fake News, slander, and so on, they will be less likely to be hurt and misled. However, since Fake News and slander and so on do not occur only on the Internet, there is a danger that users will be unable to make informed decisions when they encounter such information outside of the Internet. We believe that it is vital to identify the criteria that users want for content moderation. The following research questions were set:

- What kind of content moderation do users expect for offending content from digital platforms?
- How do users balance freedom of expression in an environment in which they do not see offending content?

Several analyses can be conducted using this research question. For example, opinions may differ depending on basic attributes such as gender and age. Alternatively, content creators on social media can be considered to have more open standards.

To clarify user perspectives, we surveyed the content moderation standards desired by digital platform users. We analyzed YouTube users, who have the largest number of viewers in the Mitsubishi Research Institute report, as a representative example of an SNS. As a preliminary step in the research design, we extracted and categorized the actions YouTube has dealt with in its Google Transparency Report. When we categorized the reasons for dealing with deleted channels, videos, and comments, we found 12 reasons for dealing with them. We selected relevant examples from specific violations exemplified in the Communication Guidelines. No specific examples were found for malicious expressions, harassing behaviors, or others. Thus, it can be inferred that “multiple policy violations” are a combination of other items. There are multiple specific examples of the remaining nine reasons.

In this study, we assumed that users would prefer different methods based on the types of violations indicated on YouTube. Therefore, in our survey, we analyzed the desired coping strategies for each violation. We also assumed that the wider the range of interpretations, the less accurate the answers. Therefore, to provide respondents with a more concrete picture of the content, we provided them with concrete examples in the questionnaire to make it easier for them to assume the title. Based on these results, we conducted a survey on nine types of violations (Table 1). Note that no consideration was given to whether the specific examples presented were representative of the number of banned content items addressed in the Google Transparency Report.

Table 1. 9 Type of violations

Q1	spoofing
Q2	Spam, Misleading expressions
Q3	Nudity or sexual expression
Q4	Child Safety
Q5	Hate speech, insulting comments
Q6	Harassment, cyberbullying
Q7	Violent or graphic language
Q8	Harmful or dangerous behavior
Q9	Promoting violence or violent extremism

Source: Created by the author based on the Google Transparency Report and Google Communication Guidelines

We made responses were developed from the following perspectives.

- Can I see the content?
- If not visible, who should do content moderation?

As a result, the following five types were prepared. We also presented respondents with the five coping strategies shown in Table 2 and selected their preferred coping strategies for each question.

Table 2. Countermeasures presented to respondents

1. If it’s not illegal, YouTube should treat this content as normal content because free expression should be protected.
2. I think there is a problem with the content, but I don’t want YouTube to deal with it.
3. It may exist on YouTube, but I would like it to be invisible to me using YouTube’s functions.
4. YouTube should not display content to anyone.
5. I don’t know what to do because I can’t imagine what kind of content it is.

Source: Created by the authors

The first choice is that they do not want to be constrained by anyone if it is not illegal in the first place. If this is the most common response, then the discussion on content moderation should only be about whether it is illegal or not. The second choice is that they acknowledge that harmful content is problematic, but they want it to be shown to themselves. This is an answer that emphasizes the right to know. The third response is that they acknowledge that harmful content is problematic, but they do not want to see it, even if others do. they think the balance of freedom of expression and the right to know allows for the distribution of content itself, however, they want to use Youtube's functions to reject that content. The fourth choice is that harmful content is a problem and should not be shown to anyone, including yourself. This is a choice that supports that digital platforms should control everything. The last choice is selected when these do not apply or when they cannot answer the question.

A web-based survey was conducted to answer these questions. A summary of the survey is shown in Table 3, and the basic attributes of the respondents are shown in Table 4 and Figure 1.

Table 3. Survey overview

Research method	Web research through crowdsourcing
Survey date and time	2023/08/04 - 2023/08/05
Number of collections	1,069 items

Source: Created by the authors.

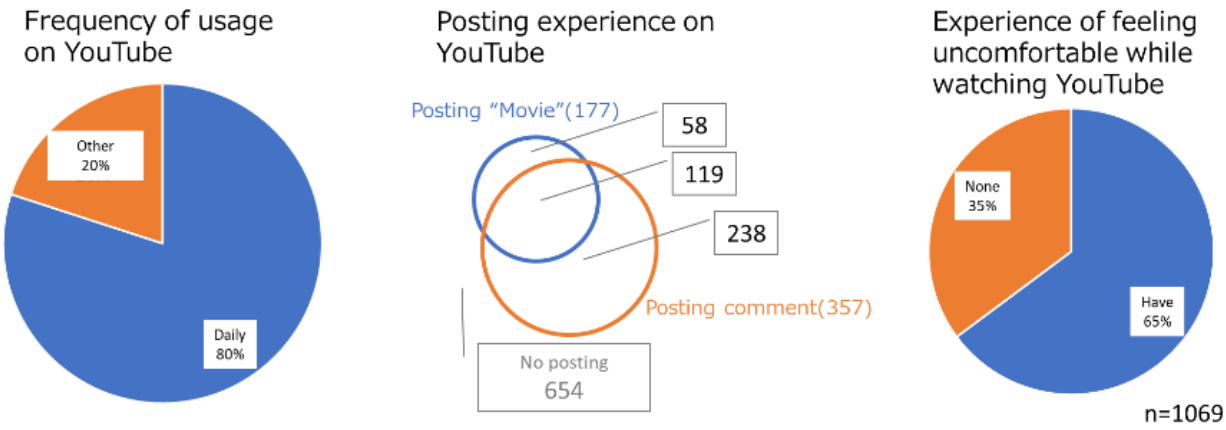
Table 4. Basic attributes of respondents

Age	Composition ratio
teenager, 20s	19%
30s	32%
40s	31%
Over 50s	18%

Gender	Composition ratio
Male	45%
Female	54%
don’t answer	1%

Source: Created by the authors.

Figure 1. Experience of respondents



4. RESULTS

Figure 2 presents the questionnaire results.

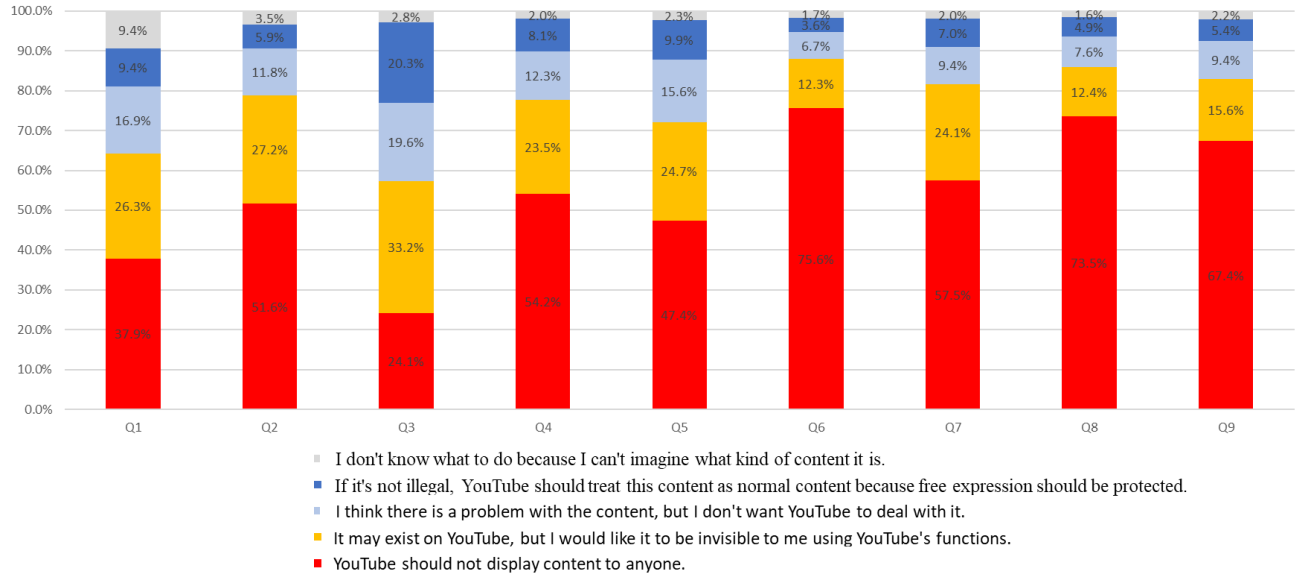


Figure 2. Questionnaire results by question

Based on the research queries of this study, we analyzed the responses and found the following:

- The number of people who believe that offending content should be freely distributed because it is not illegal (choose option 1) is small, at less than 10% for eight out of the nine questions.
- For six out of nine questions, more than half of the respondents said that YouTube should not display content to anyone.(choose option 4)
- While 24.1%, or less than one-fourth of the respondents, wished that nudity and sexual expressions were hidden (choose option 4) in Q3, 75.6%, or more than three-fourths, wished that harassment and cyberbullying were hidden (choose option 4) in Q6. Thus, there were significant differences between the questions.

In other words, for each violation area, users expect more content moderation from digital platforms, whereas there are areas where they do not want to be involved.

Next, we conducted a binomial logistic analysis to ascertain the balance between the environment in which people did not view the offending content and their freedom of expression.

The procedure is as follows. First, in the "Measures proposed to respondents," we excluded those who responded, "I don't know what to do because I can't imagine what kind of content it is." Second, "If it's not illegal, YouTube should treat this content as normal content because free expression should be protected" and "I think there is a problem with the content, but I do not want YouTube to deal with it." were defined as groups that value freedom of expression. We named this group "KEEP." "YouTube should not display content to anyone" and "It may exist on YouTube, but I would like it to be invisible to me using YouTube's functions." These are defined as groups in which the value emphasizes an environment in which the offending content is not seen. We named this group "DELETE." Third, in "Counters presented to respondents," "DELETE" is set to "1," and "KEEP" is set to "0." Fourth, we used dummy flags for the respondents' attributes and experiences obtained from the survey. Finally, we

performed a binomial logistic analysis with “Countermeasures presented to respondents” as the objective function.

All nine questions were analyzed, but only Q3 and Q6 are discussed in detail due to space limitations. The results of the study are shown in Table 5 below.

Table 5. Result of binomial logistic analysis

	Q3 N=1039			Q6 N=1051		
	Estimate	Pr(>|t|)		Estimate	Pr(>|t|)	
(Intercept)	0.371	0.000	***	0.822	0.000	***
Viewing frequency. Daily (dummy)	-0.025	0.506		-0.006	0.803	
Experience posting videos(dummy)	-0.024	0.556		-0.023	0.401	
Experience posting comments(dummy)	-0.022	0.493		0.013	0.535	
Experiencing discomfort on sns(dummy)	0.103	0.001	***	0.057	0.005	**
Age.10.20s(dummy)	-0.118	0.015	*	-0.051	0.120	
Age.30s(dummy)	-0.013	0.752		0.006	0.843	
Age.40s(dummy)	-0.051	0.225		0.010	0.727	
female(dummy)	0.327	0.000	***	0.038	0.062	
parenthood(dummy)	0.050	0.115		0.013	0.550	
Preferred Political Party. Opposition Party(dummy)	-0.021	0.609		0.039	0.165	
Preferred Political Party. Non-Party(dummy)	0.074	0.053		0.016	0.526	

Note: *** indicates 0.1%, ** indicates 1%, and * indicates significant at the 5% level

Several points can be learned from these results: The number of validated analytes was 1,039 for Q3 and 1051 for Q6.

For both questions, we found no significant differences in “Experience posting videos” and “Experience posting comments.” In other words, being a content creator may not have a significant effect on attitudes toward content moderation. Similarly, no significant differences were found in political party support. A significant difference was found in the “Experiencing discomfort on SNS” group. Those who experienced problems in the past were more likely to desire stricter content moderation on both issues. The “Age.10.20s(dummy)” group and the “female(dummy)” group were characterized only in Q3. The fact that most nude content is aimed at men indicates that women react negatively to it.

5. CONCLUSIONS AND ISSUES OF THIS STUDY

In summary, this study introduces the fact that much attention has been paid to content moderation and that it is often unclear what users want to do with offending content. We have therefore examined the following research questions: “What kind of content moderation do users want from digital platforms?” and “How do users think about the balance between an environment where they do not see offensive content and freedom of expression?”

In this study, a web-based user survey was conducted. According to the results, users’ responses to what digital platforms have defined as violations, “YouTube should not display content to anyone,” varied widely from 24.1% to 75.6%, depending on the question. We think we should separate the discussion between anti-harassment measures against which the majority of respondents want to take action and anti-nudity measures against which the majority of respondents do not want to take action.

At this juncture, we raise the following issue. This study is based on Google’s Transparency Report. The definition of problematic behavior differs from that of fake news and slander, as in the Japanese government. Therefore, the analysis results conducted in this study may not be directly applicable to the Japanese government. As this study was conducted as a fixed-point observation, it is possible that biases owing to social conditions were not removed from the analysis.

Finally, this study suggests that sex, age, and other factors may have influenced responses. Therefore, we intend to conduct an ongoing user survey. By attempting to eliminate biases owing to changes over time and social conditions, we can expect to gain new insights into the content moderation desired by users and the factors that determine it. Accordingly, we would like to address these points.

REFERENCES

ITU (2022) “The State of Broadband 2022: Accelerating broadband for new realities”, p. 2

European Commission (2022) “Digital Services Act” CHAPTER I Article 3 (t).

Access Now et al. (2021) ”Santa Clara Principles 2.0”

Democratic Party (2020) “2020 Democratic Party Platform” https://www.presidency.ucsb.edu/documents/2020-democratic-party-platform

Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, Casey Fiesler (2023) “A Trade-off-centered Framework of Content Moderation”, ACM Transactions on Computer-Human Interaction, Vol. 30

Ministry of Internal Affairs and Communications (2022) “Final Report of the Study Group on Platform Services”

Yusuke Tachibana (2022) “Trends in Section 230 of the Communications Decency Act and Implications for Provider Responsibility,” Journal of the Institute of Information and Communication Engineers, Vol. 39, No. 4, pp. 119-126

Eijiro Mizutani (2022) “Content Moderation on Social Media Platforms and ‘Freedom of Expression’,” Media Communication: Bulletin of

the Institute of Media and Communication, Keio University, No. 72, pp. 27-40

Fujio Toriumi and Tatsuhiko Yamamoto (2023) “Joint Proposal ‘Toward a Healthy Speech Platform: Digital Diet Declaration ver. 2.0’”

Mitsubishi Research Institute (2022) “Survey Results on the Actual Distribution of Defamatory Information on the Internet”

Ryosuke Watanabe, Daisuke Umemoto, Satoshi Imamura (2021) “Digital platform legal issues and practices” Aokishoin.

Shinichi Yamaguchi (2015) “A study of the actual situation of internet flames and policy responses. Social impact and defamation, restrictive identity verification systems, and the state of Internet literacy education as seen from empirical analysis.” Information and Communication Policy Review, Vol. 11, pp. 52-74