Southeast Asian content moderation protocol – a concept w/ focus on Twitter T&S model

K.S. Park, Open Net Association

kyungsinpark@korea.ac.kr

Internet was on the side of democracy

- What is democracy? Equality? Liberty? Life? No. People governing themselves
- Formation of people as agency of will, why poor ppl vote for elites? – need communication with one another
- Internet – revolutionized communication – any-to-any full connectivity – no central control – no $ charged for sending or receiving data
- Net neutrality + intermediary liability safe harbor – removing both middlemen in 2 ways – Alice's wonderland
- 2011-12: Jasmin revolution, Internet for Nobel Peace prize
- President Roh (South Korea) and President Obama(US), the first presidents "elected by the internet".

Dark side learned the internet

- Trump, source of fake news, began using the internet
- ISIS – recruiting through Twitter
- American right-wing militia – recruiting through websites
- Government disinformation campaigns during elections
- Ruling majorities using internet to spread fake news about minorities
- Ruling majorities sharing information about targets for persecution

What to do now?

- Should the internet remain empty platforms that anyone can use.
- Article 19 freedom of expression
- Article 20 ban hate speech
- But there are many speeches that are protected by Article 19 that do not violate Article 20 but harm democracy
- Can we make the internet to **take the side of democracy?**
- Do we need control to “platforms”? Let them take out bad but lawful speech?
- What to do with intermediary liability safe harbor? Are we bringing back “middleman?
  - intermediary’s right not to be associated with certain messages
  - Difference between LAW vs ETHICS → a diversity of platforms allowed to innovate with different business models

Southeast Asia: protect both rights and democracy

- Authoritarianism is rising in the SEA region, and there must be ways to maximize internet freedom so that the people can push back against the rising authoritarianism.
- **Online administrative censorship** is on the rise in Southeast Asia. Malaysia, Viet Nam, Thailand, and Indonesia implemented mandatory "notice-and-takedown" systems where criminal/civil liability are imposed on intermediaries for failure to take down or block websites when government agencies make the requests.
- **Criminalization of speech: ask them push back on non-judicial surveillance or data demands to protect anonymous speech**
- **Protecting Democracy:** Social media trolls, often aligned with governments, or often government disinformation are attacking dissident groups, human rights defenders, and vulnerable groups to shrink the latter's freedom of speech and civic space, making it difficult to achieve substantive democracy as opposed to formal democracy.

Southeast Asian Content Moderation protocol

Procedure

- How to select trusted CSOs?
- How, when, where to consult with them?
- Consultation on postings v. consultation on rules
- How to refresh groups?

Standard

- Hate speech + (include non-violent hate-mongering)
- Hate speech – (exclude minority's protest against hate)
- State actors' disinformation
- Ruling majorities' disinformation/hate speech against minority (non-protected group)

Rabat Plan of Action: 100% harm-based standard

- context
- intent
- the status of the speaker,
- content
- the reach of the speech
- and the likelihood or imminence of harm

Option: Trust and Safety Council Model

- Selection of consultants (about 40 orgs & 70 ppl, casting a wide net)
- On-going periodic update on changes in community guidelines
- Cycle: Changes – (experiment) – **T&S consultation** – public comment – (results analysis) – feedback – (start again?)
- Crisis response hotline (3/27/20), experiment (5/5/20) shared only with T&S members
- Members listed publicly and updated regularly
- In-person annual conference where more private ”products” can be shared

Strengthening our Trust and Safety Council

Friday, 13 December 2019

In 2016, we established the Twitter Trust and Safety Council, which brings together more than 40 experts and organizations to help advise us as we develop our products, programs and the Twitter Rules. We've been discussing internally and with the current members how we can make sure the Council best serves the people who use Twitter, and heard feedback that we needed to broaden membership to include a more diverse range of voices and organize members to have deeper conversations.

Going forward, the Council will be made up of several groups, each focused on advising us on important issues that contribute to real-world

”Serving the Public Conversation”

- 5/10 and 5/11 T&S Council Office Hours
- Less than 1% of accounts make up the majority of accounts reported for abuse, but a lot of what’s reported does not violate our rules. While still a small overall number, these accounts have a disproportionately large – and negative – impact on people’s experience on Twitter. We want to be proactive in addressing disruptive behaviors that negatively impact the health of conversations. (5/16/18)
- A few examples of signals we are integrating include (5/9/18) :
  - If an account has not confirmed an email address
  - If the same person signs up for multiple accounts simultaneously
  - Accounts that repeatedly Tweet and mention accounts they do not follow
  - Behaviors that indicate a coordinated attack.
- This content will remain on Twitter, and will be available if you click on “Show more replies” or choose to see everything in your Search setting. (5/16/18)

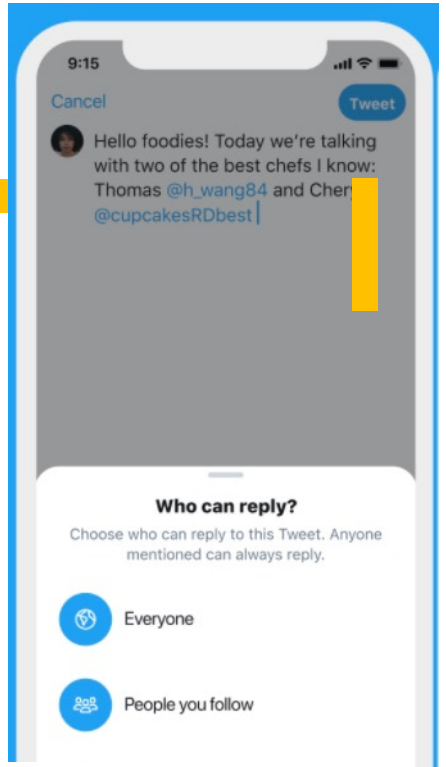
”Hateful conduct policy”

[Religious Group] should be punished. We are not doing enough to rid us of those filthy animals.

- We will be updating our Hateful Conduct policy to include content that dehumanizes others based on protected category or membership in an identifiable group. Research shows that this speech has the potential to normalize offline violence, and we want to mitigate the potential impact. This behavior is currently in violation of our policies when it is targeted at someone else (e.g, @mention, #name, tagged in a photo, etc). We will be expanding our policy to include content that is not explicitly targeted at an individual. (9/20/18)
- we heard compelling rationale that ‘all identifiable groups’ is too broad a category and we must address additional factors before we can include language directed at other protected groups.. . we’ve scaled back the originally proposed scope of the policy (7/10/19)
- **Was I wrong?** E.g., Korea’s Sewol ferry victims and Itaewon stampede victims

“Author Moderated Replies” and Author-chosen Reply settings

- Giving back control to the users (I: Creating local safe civic spaces) (December 2018)
- To strike a balance between giving authors control and maintaining transparency, Tweets that are hidden will still be viewable. This transparency was highlighted by council members and we are grateful for your feedback. (11/22/19)
- Authors can choose who can reply – Started with a confidential experiment (5/19/20) → sharing result of experiment (8/8/20):



Since we started testing the feature, we've learned that **conversation settings can help people feel safer and more comfortable Tweeting**. Here is what people shared:

- People feel protected from spam and abuse. People who face and report abuse are 3x more likely to use these settings.
- Tweets using these settings receive fewer potentially abusive replies. And people aren't finding a new way to reply through unwanted Retweets with Comments or Direct Messages.
- The feature is a new method to block out noise. 60% of people who used this during the test didn't use mute or block.
- People are sharing more thoughts on politics and social issues. Tweets using these settings about Black Lives Matter and COVID-19 are on average longer than those that don't use t

Political advertising

- **political content** and **cause-based advertising**. The development of each policy was guided by a set of beliefs:
- Political message: reach should be earned, not bought.
- Advertising should not be used to drive political, judicial, legislative, or regulatory outcomes; however, cause-based advertising can facilitate public conversation around important topics.
- Advertising that uses micro-targeting presents entirely new challenges to civic discourse that are not yet fully understood. The impact on voters of tools like micro-targeting is not well understood, so we have put in place clear restrictions on how cause-based ads can be targeted. (11/22/19)

Crisis Response

- Twitter is engaging with trusted experts and partners like you to respond to the COVID-19 pandemic. If you have concerns about content that should be reviewed, potential mistakes in our automated systems, or COVID-19-related issues, please email NGOHelp@twitter.com to escalate these issues. We ask that you please not share this alias publicly. (3/27/20)
- Starting today, we're introducing new labels and warning messages that will provide additional context and information on some Tweets containing disputed or misleading information related to COVID-19. (5/11/20, no consulting but it is okay)

State actor labels: Do we need to do more?

Aggregate analysis – e.g. Vulnerable Person rule

- Often postings individually do not present likelihood of harm but pose harm together.
- For example, X postings from religious leaders may attack an abstract group of progressive reformers and Y postings 1 week later from the same leaders may just list the names of progressive reformers.
- X postings will not be moderated b/c of lack of specificity. Y postings will not be moderated b/c of no incitement. But together they send clear signals to mobs/trolls to attack.
- Twitter's T&S Council's last project was to develop "Vulnerable Person" rule, designed to deamplify the postings listing HRDs and journalists who are likely targets of trolls/mobs.
- These aggregated harmful postings will be spread over time and place. There has to be an ongoing relationship between CSOs and platform operators.

Conclusions

- Platforms should not emulate governments, i.e., retain all postings protected by international human rights standards.
- Governments come into power with a lot of promises made to voters. Much content regulation is ideology-driven and rightly so as people have democratic sovereignty to shape their collective (communicative or discursive) future.
- Platforms can be more pragmatic, i.e., harm-based Rabat-Plan-type principle. Instead of collectivistic goals, platforms can focus on preventing actual harms of speech from being realized. Also, remedies can be more nuanced than direct takedowns, e.g., deamplifying, hiding behind buttons.
- In doing so, “State actor labels”, “political advertising”, ”serving the public policy”, hate speech rule with broader protected groups, and other rules of the former Twitter are examples of harm-based rules.
- Finally, instead of posting-by-posting analysis, we need a group posting analysis that measure the likelihood of harm from an aggregate of postings, which requires a standing relationship with CSOs through the Trust-and-Safety Council-type congregation.