Combating hate speech on social media platforms

“2020 Conference on Combating Hate and Discrimination– Implementing Strategy and Plans of Action to Address Hate and Discrimination”

(Sept. 17, 2020)

Kim Min Jeong

(Professor, Media Communication Division, Hankuk University of Foreign Studies)

Overview

- Intro.
- National regulation, overseas references, implications

  (1) Government regulations

  (2) Regulation by individual platform

  (3) Self-regulation by Korea Internet Self-governance Organization(KISO)
- Summary and additional proposal

Intro.

1) As for regulating content on the internet, it is desirable to minimize government regulation and vitalize self-regulation
   - Constitutional Court Decision 99Hunma480(June 27, 2002): Unconstitutionality of a provision on rebellious communication of 「Act on Telecommunications Business」 *cf. Reno v. ACLU* (1996)
   - Korea has bigger impact of administrative regulations than other countries regarding regulation of content on the internet, while it has insufficient history and experience of self-regulations

2) A need to re-examine 20th century's methods of regulating expression in consideration of changed media environment and harmful impact of internet expressions

Intro. (cont.)

3) Vast amount of expressions subject to regulation is the biggest challenge that makes it difficult to regulate expressions on a large-scale social media platform
   - Moderation by the methods of legacy media is impossible; It is difficult to apply a method where individual online community decides and discusses a rule
   - For a large-scale platform, managing posts and comments is industrial work, not artisanal work (Gillespie, 2018, p. 77).
   - A method which combines editorial review+ management based on report of users+ automatic detection and deletion using technologies

Domestic regulation (1) Government regulation

- No legislation directly prohibits hate speech
- Application to hate expression that constitute defamation or insult if possible. However, there is a limitation that it is possible only when individual victim is specified.
- 「Act on Promotion of Information and Communications Network Utilization and Information Protection, etc.」('Act on Information and Communication Network'):
  - " illegal information ", " harmful information(harmful information for adolescents)", information that violates one's personal rights(Temporary measure against defamation and violation of privacy) → Partially regulates hate speech(when an individual is specified)
- Unsound information("Information that undermines sound communication ethic") is subject to administrative review. 「Review regulation on information and communication」 includes regulations that can be used to regulate online hate expression when an individual victim is not specified.
  - Article 8.3 (f) prescribes "content that discriminated against or promotes discrimination based on sex, religion, disabilities, age, social status, origin, race, region or occupation without reasonable grounds" as a type of information subject to review (Review on 'information that undermines social integration and social order")

Overseas cases

- Germany: 「An act to improve law enforcement on social networks(Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken)」('Network law enforcement act') is in force since Jan. 1, 2018.
  - It is obligated to promptly delete or prohibit access(within 24 hours for explicit illegal posts and within 7 days for other illegal posts) upon user report regarding 21 types of illegal information(6 types are regarding hate expression) prescribed in German criminal code.
- France: 「An act to respond to hate contents on the internet (loi visant à lutter contre les contenus haineux sur internet 」)(Avia act) was passed on May 13, 2020.
  - Hate expression prescribed as illegal under the Freedom of Media Act of France(Article 24(promotion of discrimination, hostility, violence) and article 33(crime of collective insult) should be deleted or its access has to be prohibited within 24 hours upon user report.(Cf. Deletion or prohibition of access within 1 hour for contents related to terror or child sexual exploitation.
  - Applies to a platform operator with certain monthly access volume in France

Implication

- Korea already has a clause in Act on Information and Communications Network that obligates online service providers to play a certain role to prevent distribution of illegal information. Treatment based on user report under legislations of Germany and France is similar to temporary treatment under the Act on Information and Communications Network.
- In Korea, various types of online information are already defined as illegal information and information that infringes personal rights, being subjected to deletion and temporary treatment. However, hate expression is not defined as such types of information. In addition, the number of sanctions pursuant to Review regulation on information and communication which can be applied when an individual is not specified is not higher when compared to the total number of communication review.
- If Korea wants to enact a legislation to effectively regulate hate expression on social media platform, we need to go through an order that is opposite to the way taken by Germany and France. That is, hate expression itself has to be defined as illegal information or information that infringes personal right.

Domestic regulation (2)
Regulation by individual platform

- Foreign operator:
  - Facebook's Community Standards on hate expression & Youtube's Community Guidelines on Contents with hatred: Disclose standard, example and enforcement status
  - As universal application of a single policy, there is no special consideration for Korea
  - According to a report of IT related media of the U.S., Facebook has 15,000 content moderators who monitor postings that violate Facebook's Community Standards as of early 2019. The number of personnel in charge of reviewing policy violation of Korean users' posting is unknown.
  - After COVID-19 outbreak in March 2020, the number of content moderator is decreased. The ratio of AI playing a role of final decision maker is increasing: Youtube focuses on excessive regulation(platform's responsibility) and Instagram by Facebook focuses on minimum regulation(freedom of expression) when operating algorithm(Report of ZDNet Korea on August 26, 2020)

Domestic regulation (2)
Regulation by individual platform(cont.)

- Domestic operator:
  - Naver and Kakao briefly deal with hate expression in their posting and comment operation policies
  - Personnel attack through comments on news related to celebrities and athletes has become a serious issue → Recently, comment function on entertainment news/sports news is abolished
  - Since Feb. 2020, Kakao added “discrimination and hate” in news comment reporting items and operates a policy to hide or delete comments that have been reported as hate expression

Reference regarding regulation by individual platform: Company's content moderation principle

- Content Moderation principle of internet platform business, David Kaye, UN Special Rapporteur on Freedom of Expression(Oct. 2019)

1) Legality:
   - Clearly express specific scope of hate speech and company's policy
   - Transparency and consistency of execution
   - Judgment on detailed context should be made by a person because it cannot depend on artificial intelligence or automatic technology, and the judgement by a person should be based on content acquired from community's experience that actually experienced hate speech
   - Researches show that regulation of hate speech through automated filtering technology is more harmful to groups that have been underrepresented historically
   - Big companies have a responsibility to invest recourses and to share their knowledge and technology as open source to enable small businesses to access those technologies

Company's content moderation principle(cont.)

2) Necessity and Proportionality:

- Consider measures other than posting deletion:
  - Indicate original source of posting; develop a rating system that limits usage of posting; a measure for temporary limitation during post review period; limit pursuit of economic benefit by speakers of hate speech; warning; blocking; a measure to minimize spread of posting; a measure to prevent usage of bot or systematized public action; restrictive measure per region; a measure to encourage alternative expression
- Necessity and proportionality should be openly explained upon request of a person affected by platform's measure

Company's content moderation principle(cont.)

3) Devise a measure for remedy:

- Prepare a measure against excessive reports misusing a provision on prohibition of hate expression
- Create a transparent and easy-accessible process to enable users to present an objection regarding platform business's decision
- Operator should make the following efforts:
  - Education on dangers of hate expression and the way hate expression silence social minorities; Visualization of responsive measures against hate speech; public denunciation of hate expression through public campaign or public figure's speech; and devise a measure to evaluate scope of the issue and effectively prevent spread of hate expression through cooperation of social scientist

Domestic regulation (3) Self-regulation by KISO

- Korea Internet Self-governance Organization (KISO)
  - Established in 2009, Participation of 12 Information and communication service provides such as Naver and Kakao
  - Declared core principle of self-regulation through KISO policy and regulate contents such as posting, comments and search word
  - Aims for a “voluntary self-regulation” model without government intervention(KISO, 2018, p.28)
  - KISO Policy Committee, composed apart from Board of Directors, has 5 outer commissioners and 5 commissioners from member companies. Outer commissioners are legal experts and scholars.

Domestic regulation (3) Self-regulation by KISO

- KISO policy regarding hate expression:

1) Policy regarding search word: When a related search word contains “a word that disparages certain region, religion, belief, disability, race or country of origin and thus there is a possibility that exposure of such word may lead to excessive social conflict”, a member company may delete or exclude that search word upon user request(Article 13.2, paragraph 2(3)) -> policy on related search word to mitigate social conflict
2) Policy regarding posting: Member companies may delete or take necessary measures regarding “a posting that uses insulting or aversive expression against certain group that can be divided by region, disability, race, country of origin, sex, age or occupation and thus causes significant insult or disadvantage against that particular group”(Article 21). -> policy to mitigate discriminative expression

Application of KISO policy regarding hate expression

- As of early April, among 193 KISO review decision disclosed online, ‘policy on related search word to mitigate social conflict’ is applied to 3 cases, and ‘policy to mitigate discriminative expression’ is applied to 4 cased.
- ’Pinko’ is the only related search word decided to be deleted
- Deleted posting include:
  - A posting that contains racial discrimination by using expression such as ‘dog’ and ‘Xiongnu’ or indecent or insulting images
  - A posting that derogates Jeolla region by connecting it to pro-North Korea without grounds

Application of KISO policy regarding hate expression (cont.)

- Posting decided as 'not applicable'(not deleted)
  - Among postings with regional discrimination, a case where an expression of 'fraud' was used or a posting with abstract of newspaper or historical material was not deleted because it is difficult to be concluded as insulting or aversive despite insufficient rationality and extreme inclination.
  - Hate expression constitutes discrimination against sexual minorities was not deleted because that expression does not include a "word or symbol" that is discriminatory or aversive.
- That is, in operating the provision concerned, not only whether it is an expression that justifies, promotes and reinforces discrimination but also whether the expression itself is aversive or insulting is an important factor considered.

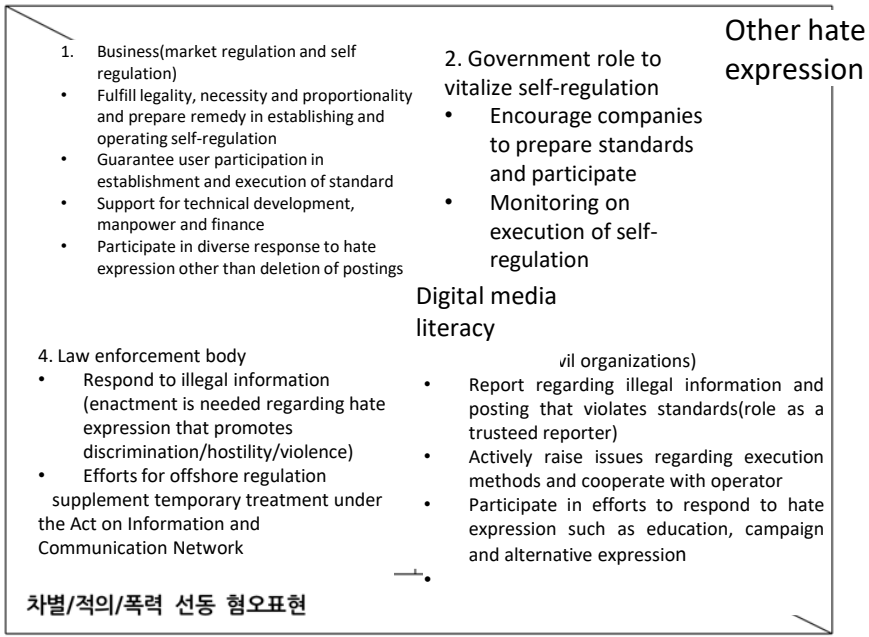
Foreign reference regarding self-regulation

- ‘Code of Conduct on Countering Illegal Hate Speech Online’
  - Jointly adopted by Facebook, Microsoft, Twitter and Youtube upon request of the EU
  - 12 practices: Include measures to enable participation of not only companies but also government and users as main agent of self-regulation
  - Main content: Each company deletes or blocks access to hate expressions that are prescribed as illegal by EU member states upon user reports, and each operator devises a provision regarding hate expression in its community guideline, assigns a team to review user reports and prepares an effective process to review reports.
- EU’s Commissioner on for Justice, Consumer and Equality periodically evaluates company’s implementation status of code of practice based on result submitted and issues a report.

Implication regarding Self-regulation

- It is suggested to reflect EU'S measures to enhance participation of users and government to operation of domestic self-regulation
- It is needed to cooperate with civil society experts to divide hate expression that promotes violence and act of hatred into different ratings, support civil society organizations to enable them act as a trusted reporter/flagger, enhance cooperation with them to provide education to respond to hate expression, and support campaign for alternative expression.
- Critic that regulation on hate expression by individual operators or self-regulation body is implemented in an arbitrary and discriminatory way. In order to address this criticism, it is needed to expand participation of user including civil society organization in implementing self-regulation.

Conclusion: Hate speech regulation model on social media platform



hate expression that promotes discrimination/hostility/violence

Additional proposal

- Hate speech on social media platform is not an individual and independent issue and thus it cannot be solved without comprehensive response by various agents both online and offline.
  - Prompt enactment of a comprehensive anti-discrimination act is needed.
  - Hate speech on the internet is likely to be linked with cyber bullying or stalking, and it may lead to disclosure and dissemination of victim's personal information when it is carried out in real-time. Revision of related legislation and creation of countermeasures is urgent.
  - It is needed to look at hate expression in media report and seek for improvement measures. The media should be aware of this issue and make efforts for improvement as well.

*Content of this presentation is based on “Combating hate speech on social media platform”, <Media Culture Research>, Kim Min Jeong(2020), 32(1), 7-54