The Control of Hate Speech on Social Media: Lessons from Sri Lanka

POLICY ISSUE

As hate speech on social media becomes an ever-increasing problem, policymakers may look to more authoritarian measures for policing content. Several countries have already, at some stage, banned networks such as Facebook and Twitter (Liebelson, 2017). The government of Sri Lanka recently enacted a similar measure: a complete shutdown of Facebook, WhatsApp, Viber and Instagram on the basis that Facebook was used to spread messages of racial violence.

However, analysis reveals these blocks to be ineffective (Wijeratne, 2018). Based on discussions with Facebook, this brief presents the more technically challenging approach of detecting and legally limiting hate speech as a more effective approach.

While this is based on Sri Lankan context, we extract lessons for countries across the global south to propose a role that governments can fill to more effectively curb the spread of hate speech.

TARGET POLICYMAKER/S

ICT bodies maintained by or affiliated with national governments.

RECOMMENDATIONS

Build a public, national corpus of written content in all major languages in the nation and make it available in machine-readable formats.

Build and share indexes of slurs and racial epithets and the context around them, which can be shared to Facebook for use in their hate speech detection technologies.

To aid human content moderation, encourage Facebook to hire a minimum number of content moderators fluent in national languages, such that native speakers will be present to moderate content.

Compare and contrast Facebook Community Standards with local hate speech laws and seek region-specific amendments or expansion.

Strengthen legal mechanisms through which local justice can be implemented on the perpetrators of hate speech that may escape Facebook's definitions and detection mechanisms.

THE RESEARCH

THE FOLLY OF BLOCKING CONTENT

On the 6th of March, the government of Sri Lanka declared a State of Emergency following racially instigated attacks against the Muslim population of a town in Kandy (Mashal, Bastians, 2018). The government also enacted a block on Facebook and related platforms (Mallawarachi, 2018; Wijeratne, 2018). Government spokespeople laid the blame on Facebook for enabling the spread of hate speech regarding the incident (Bengali, 2018; Rajagopalan, Nazim, 2018).

However, analysis of the posting times of over 60,000 posts reveals this measure to be technically ineffective (Wijeratne, 2018).

Furthermore, the conversation between journalists, activists and media spokespeople on Twitter, which was left unblocked, suggested loss of political goodwill, fear regarding both the political control of social media (Wijeratne, 2018) and doubts about effectiveness of both Facebook to enforce its own policies (CPA, 2018).

These are outcomes posited in The Internet Society Perspectives on Internet Content Blocking: An Overview (Seidler, Robachevsky, 2017). The paper in question describes five possible ways in which to enact a block based on public policy considerations, and notes that these methods **don't solve the problem**, as sufficiently motivated users can circumvent them with ease; they also **cause collateral damage**, as precise targeting is difficult; they **put users at risk**, by forcing even law-abiding citizens to take risky, alternate routes to access their communities; they **also drive services underground,** and away from the purview of both the platform and from law enforcement. Some of these user reactions were indeed observed in Sri Lanka, as Google searches for the keyword "VPN" - denoting a popular method of circumventing website blocks- rose dramatically across the island right from the start of the period of blocking.

This situation is not unique to Sri Lanka: similar incidents of hate speech, and similar user reactions have been making headlines across the world, notably in Myanmar (Taub, Fischer, 2018; Gillbert, 2018).

Perhaps the only state that can be said to have successfully implemented content blocking measures is China, as documented by Barme and Ye (1997), Zittrain and Edelman (2003), and Denyer (2016). It should be noted, however, that the efficacy is still technically questionable, and the Chinese implementation hinges on both powerful technical infrastructure, intimidation, authoritarian policies and networks of power brought on by the massive userbase the Chinese state controls (Waddell, 2016). The concluding argument, drawn from the Internet Society, is that:

Where there is wide-ranging agreement on illegal content, the best solution to the problem is removal of the content at the source.

THE DIFFICULTIES OF REMOVING HATE SPEECH AT THE SOURCE

Facebook detects and removes hate speech on their platform primarily via two methods: when users of the platform report it, and via detection by their machine learning tools (Chen, 2014; Terdiman, 2018). Even so, as Sri Lanka demonstrated, these have not been entirely successful.

We posit that these are a result of three components: firstly the framework under which these efforts operate: the Facebook Community Standards (Facebook, 2018), a systematic framework for classifying hate speech under various criteria, with provisions for slurs, epithets, speech against protected groups, and various mechanisms for classifying the severity of a threat.

Due to the lack of a universal definition of hate speech, there are mismatches between this framework and national law. As a case in point, the ICCPR Act 56 of 2007 upheld by the Sri Lankan government mandates a non-bailable sentence with a maximum of ten years' imprisonment for any propagation of war. Facebook's Community Standards (2018) suggest that general calls to war, lacking specifics are of relatively low threat levels. This, and other parameters of Facebook policy (Propublica, 2017) prevent Facebook, even working at maximum capacity, from being a silver bullet solution to hate speech.

Thus, it is paramount that policymakers both understand the guidelines and work with Facebook and law enforcement to fill in region-specific gaps that such an international framework may be unable to resolve.

Secondly, we come to the human component: a lack of moderators fluent in a given language - or moderators unaware of the context in which a word might be deployed - lead to repercussions at national scales (Wong, 2017; Bengali, 2018).

Thirdly, we come to Facebook's hate speech detection via machine learning. One of the key requirements of effectively hate speech analysis are language resources. Much of the Global South is at a disadvantage here: most languages we use lack the digital lexicon required for computational analysis (Nakov and Ng, 2019; Wang, Nakov and Ng, 2016) - that is to say, they. Thus, if Facebook does not have the basic building blocks to analyze content in our languages, they cannot impose scalable control of hate speech. However, given that Facebook is actively making investments in solving this issue, both on a technological and an educational fronts (Daily Mirror, 2018), it is our recommendation that policymakers seek to understand the challenges faced by platforms in implementing these measures and seek to aid these for regions under their purview.

REFERENCES

Mashal, M., & Bastians, D. (2018, March 06). Sri Lanka Declares State of Emergency After Mob Attacks on Muslims. Retrieved from https://www.nytimes.com/2018/03/06/world/asia/sri-lanka-anti-muslim-violence.html

Mallawarachi, B. (2018, March 07). Sri Lanka blocks social media as anti-Muslim rioting flares. Retrieved from https://www.apnews.com/f9ed5422ead44a97a40399c01771dad3

Wijeratne, Y. (2018, March 19). The March 2018 Social Media Block: A 30,000 foot view. Retrieved from https://drive.google.com/file/d/1PcCLYh20K2a73iPGwmvub-Ya16lwTQpL/view

Bengali, S. (2018, March 29). Muslims faced hatred and violence in Sri Lanka. Then Facebook came along and made things worse. Retrieved from http://www.latimes.com/world/asia/la-fg-srilanka-facebook-20180329-story.html

Rajagopalan, M., & Nazim, A. (2018, April 7). "We Had To Stop Facebook": When Anti-Muslim Violence Goes Viral. Retrieved from https://www.buzzfeed.com/meghara/we-had-to-stop-facebook-when-anti-muslim-violence-goes-viral

Taub, A., & Fisher, M. (2018, April 21). Where Countries Are Tinderboxes and Facebook Is a Match. Retrieved from https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html

Gillbert, D. (2018, April 11). Hate speech is still going viral on Facebook in Myanmar, despite Zuckerberg's promises. Retrieved from https://news.vice.com/en_ca/article/8xdw83/zuckerberg-

says-facebook-is-taking-its-myanmar-problem-seriously-activists-say-thats-bs

Liebelson, D. (2017, June 24). MAP: Here are the countries that block Facebook, Twitter, and YouTube. Retrieved from https://www.motherjones.com/politics/2014/03/turkey-facebook-youtube-twitter-blocked/

Center For Policy Alternatives (2018, April 11). Open letter to Facebook: Implement Your Own Community Standards. Retrieved from http://www.cpalanka.org/open-letter-to-facebook-implement-your-own-community-standards/

Seidler, N., & Robachevsky, A. (2017, March 24). An Overview of Internet Content Blocking - Internet Society. Retrieved from https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/

Baume, G. R., & Ye, S. (1997, January 6). The Great Firewall of China. Retrieved from https://www.wired.com/1997/06/china-3/

Zittrain, J., & Edelman, B. (2003). Internet filtering in china. *IEEE Internet Computing,7*(2), 70-77.

Denyer, S. (2016, May 23). China's scary lesson to the world: Censoring the Internet works. Retrieved from https://www.washingtonpost.com/world/asia_pacific/chinas-scary-lesson-to-the-world-censoring-the-internet-works/2016/05/23/413afe78-fff3-11e5-8bb1-f124a43f84dc_story.html?utm_term=.123d476974b7

Waddell, K. (2016, January 19). Why Google Quit China-and Why It's Heading Back. Retrieved from https://www.theatlantic.com/technology/archive/2016/01/why-google-quit-china-and-why-its-heading-back/424482/

Chen, A. (2014, October 23). The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. Retrieved from https://www.wired.com/2014/10/content-moderation/

Terdiman, D. (2018, May 02). Heres How Facebook Uses AI To Detect Many Kinds Of Bad Content. Retrieved from https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content

Wong, K. (2017, April 26). The unexpected origins of the controversial Myanmar word 'kalar' | Coconuts Yangon. Retrieved from https://coconuts.co/yangon/features/kenneth-wong-the-origins-of-the-controversial-myanmar-word-kalar/

Nakov, P., & Ng, H. T. (2009). Improved statistical machine translation for resource-poor languages using related resource-rich languages. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 - EMNLP 09*.

Wang, P., Nakov, P., & Ng, H. T. (2016). Source Language Adaptation Approaches for Resource-Poor Machine Translation. *Computational Linguistics,42*(2), 277-306.

Facebook launches initiative to improve Digital Literacy in Sri Lanka. (2018, May 18). Retrieved from http://www.dailymirror.lk/article/Facebook-launches-initiative-to-improve-Digital-Literacy-in-Sri-Lanka-150103.html

Community Standards | Facebook. (2018). Retrieved from https://www.facebook.com/communitystandards/

Angwin, J., & Grassegger, H. (n.d.). Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. Retrieved from https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

ACT, No. 56 OF 2007. Ministry of Justice, Sri Lanka. Retrieved from https://www.lawnet.gov.lk/2016/12/07/act-no-56-of-2007/

Yudhanjaya Wijeratne | LIRNE*asia* | 12, Balcombe Place, Colombo 08, Sri Lanka | yudhanjaya@lirneasia.net.

This work was carried out with financial support from the International Development Research Centre, Canada. The views expressed in this work are those of the creators and do not necessarily represent those of the International Development Research Centre, Canada or its Board of Governors.