Nesta attacks study

Patrycja Tempska and Rafal Urbaniak

Contents

1 Section Abstract

This article describes an experimental intervention study based in a naturalistic, digital setting (Q&A forum - Reddit), utilizing a collective intelligence approach to content moderation and reduction of the level of verbal aggression among a selected group of Reddit users who regularly attack other community members. Collective Intelligence in this sense means exploring the collaboration between human and machine intelligence to develop solutions to social challenges. Artificial Intelligence was used to detect verbal aggression (personal attacks) and notify human volunteers about attacks. Volunteers after receiving notifications employed interventions based on norm or empathy promotion. We find that only those who were sanctioned with norms-inducing interventions had their personal attacks' user significantly decreased.

```
2+2 #use this formatting for chunks
```

```
## [1] 4
```

2 Section 2

Although much effort has been made in order to tackle the problem of verbal aggression and harassment online, looking at various reports and surveys, it remains a common hindrance for people engaging with social media in their everyday lives. The situation got exacerbated amidst the COVID19 pandemic, during which a majority of our social life moved to cyberspace. During this shift, there was an increase in cyberbullying attitudes and perpetration (Barlett, Simmers, Roth, & Gentile (2021)), 90% increase in public reports of illegal online content[1], including 114% increase in non-consensual sharing of intimate images, 30% increase in cyberbullying, as well as 40% of increase in adults reporting online harassment. According to a report conducted by company L1ght \footnote{\href {https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf}{https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf}}, hate speech directed towards China and the Chinese went up by 900% on Twitter. Gaming platforms were in the spotlight as well, with a 40% increase in toxicity on Discord.

But alongside the growing need for even more efficient and proactive moderation, the capacity to put it into life did not go hand in hand, forcing companies and policymakers to rethink the current model of moderation processes and workforce. Due to the COVID19 restrictions including social distancing,

[1] https://www.aspistrategist.org.au/australias-esafety-commissioner-targets-abuse-online-as-covid-19-supercharges-cyberbullying/

a lot of those serving the role of moderators had to be sent home[2] without the ability to work remotely because of the constraints affiliated with restrictive non-disclosure agreements (NDA). gerrard2020covid19

References

Barlett, C. P., Simmers, M. M., Roth, B., & Gentile, D. (2021). Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic. *The Journal of Social Psychology*, 1–11.

[2] https://qz.com/india/1976450/facebook-covid-19-lockdowns-hurt-content-moderation-algorithms/