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 Introduction

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.

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

But alongside the growing need for even more efficient and proactive moderation, the capacity to execute it 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, 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) among others. Curtailing the moderators' workforce was accompanied by more agency given to algorithms and AI-based moderation. Those changes, as argued by Gerrard (2020), can be seen as a serious red flag in terms of safety for all users on online platforms.

3 Automated versus Human-based Moderation

The hindrances and threats that go along with the Artificial Intelligence-based methods for moderation have been widely debated, with the most critical discussions revolving around technology performance (MacAvaney et al. (2019), Schmidt & Wiegand (2017)). State-of-the-art solutions are mostly governed by statistical methods including deep learning and machine learning (LeCun, Bengio, & Hinton (2015), Sejnowski (2020), Jordan & Mitchell (2015)). Their performance is inherently tied to the amount of data being fed to the system and the quality of its annotation. At different stages of the process, from datasets gathering and preparation, annotation to the training or algorithms themselves, biases seem to be omnipresent (Binns, Veale, Van Kleek, & Shadbolt (2017), Geva, Goldberg, & Berant (2019) Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2021)). Users' of online services are also creative in their strategies to circumvent automated content moderation systems and as shown by Gröndahl, Pajola, Juuti, Conti, & Asokan (2018), current techniques are vulnerable to the most common evasion attacks like word changes (inseting typos and leetspeak), word-boundary changes (inserting or removing whitespace), or word appending (appending common or non-hateful words like "love"). Generalisability of the models - an ability to perform well on datasets coming from sources other than the one used for training are an important shortcoming as well (Yin & Zubiaga (2021), Swamy, Jamatia, & Gambäck (2019), Rosa et al. (2019)). As shown by Wu, Ribeiro, Heer, & Weld (2019), Lipton & Steinhardt (2019), and Musgrave, Belongie, & Lim (2020) in practice, creations of models often lacks thorough error analysis and legitimate experimental methodology, which can result in non-reproducibility. This is also connected with a potential lack of thorough understanding of the limitations of the models and spurious conclusions being made to a wider public. Specifically, @Lipton & Steinhardt (2019) distinguishes four dysfunctional patterns occurring in the current research paradigm in the industry and academia alike. First, the inability to draw a clear distinction between speculation and explanation, with the first one often being disguised as the second. Second, inability for successful identification of the sources of empirical gains (whether it was problem formulation, optimization of the heuristics, data-preprocessing, hyperparameter tuning, or perhaps yet another aspect). Third, "mathiness" - the use obscure language and often covering weak argumentation with the alluring but often apparent depth of technical jargon. Last but not least - misuse of language. This includes suggestive definitions without proper explanation of what they mean in the context (e.g. inflating good performance in simple NLP tasks to human-level natural understanding), overloading the papers with technical terminology, or suitcase words (those words that can encompass a variety of meanings, e.g. consciousness).

Yet another obstacle in the process is the lack of gold standard in dataset creation and taxonomies of abusive language being used for instance in the process of annotating different datasets. Frequently people obtain data from various sources and do not follow any universally used instructions when it comes to its annotation, leading to discrepancies between various datasets being tagged within one domain (e.g. hate speech). Lack of expert annotators and proper annotation criteria and instructions are also widespread, with the common practice hiring untrained workers from Mechanical Turk or other crowdsourcing platforms.

Although there are some initiatives developed in response, most notably, functional tests for Hate Speech Detection Models created by Röttger et al. (2020), or the Online Safety Data Initiative (OSDI) LINK https://onlinesafetydata.blog.gov.uk/about-us/, focused on projects related to improving access to data, standardizing the description of online harms, as well as creating tools and benchmarks for evaluation of technologies focused on safety, much effort must be made before wider adoption of such solutions comes into force.

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

At the same time, only automated methods can scan through the massive amount of content being generated every day on different platforms. On Facebook, there are more than 3B comments and likes daily (https://martech.org/facebook-3-2-billion-likes-comments-every-day/), 500M tweets are sent daily on Twitter (https://www.oberlo.com/blog/twitter-statistics), and over 2B comments made by users of Reddit in 2020 (https://old.reddit.com/r/blog/comments/k967mm/reddit_in_2020/) which is almost 3M comments made daily. With this amount of content, it's either impossible or extremely costly to scale the moderation workforce. One can also have doubts about the ethical aspects of hiring workers who are often unaware of how this kind of task will affect their well-being. Being submersed in the cyber-Augean stables takes a toll on many people. As examined by Roberts (2014) & Roberts (2016), workers who are hired for such tasks are often low-status and low-wage, isolated, and asked to keep what they've seen in secret under restrictive NDAs. Screening through the reported user-generated content is connected with exposure to violent and deeply disturbing materials, with child pornography, murders, or suicides as examples of the most extreme cases. This can lead to serious psychological damage, like depression, or PTSD (Roberts (2014)). Some of the employees filed a lawsuit against Facebook and as a result, the company agreed to pay $52M in compensation for mental health issues developed during the job (https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health). Taking into consideration that Facebook employs 15K moderators (https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=6c6bdbbc54d0) and most likely more are needed to keep up with the growing amount of content, with the parallel considerations about the negative effects of content moderation on mental health, a collaboration between humans and machines in this area seems inevitable. There is yet another aspect here - what moderators deal with is mostly content reported by humans. And as shown in various studies and reports, a lot of children, teens or even adults do not report cyberbullying or harassment online (LINKS: https://www.ctvnews.ca/canada/as-the-pandemic-forces-us-online-lgbtq2s-teens-deal-with-cyberbullying-1.5430945 ; https://www.adl.org/free-to-play-2020 ; https://www.ditchthelabel.org/wp-content/uploads/2017/05/InGameAbuse.pdf).

4 Pro-active and reactive moderation

#Collective Intelligence Approach
#Experimental Design
#Results

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.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *International Conference on Social Informatics*, 405–415. Springer.

Gerrard, Y. (2020). <? covid19?> The COVID-19 mental health content moderation conundrum. *Social Media+ Society*, *6*(3), 2056305120948186.

Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. *arXiv Preprint arXiv:1908.07898*.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is" love" evading hate speech detection. *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, *17*(1), 45–77.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One*, *14*(8), e0221152.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Musgrave, K., Belongie, S., & Lim, S.-N. (2020). A metric learning reality check. *European Conference on Computer Vision*, 681–699. Springer.

Roberts, S. T. (2014). *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

Roberts, S. T. (2016). *Commercial content moderation: Digital laborers' dirty work*.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., … Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*, 333–345.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2020). Hatecheck: Functional tests for hate speech detection models. *arXiv Preprint arXiv:2012.15606*.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(48), 30033–30038.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 940–950.

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (2019). Errudite: Scalable, reproducible, and testable error analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598.