Ethical considerations in Wikipedia automatic abuse detection system

- Sameer Singh (SS8GC@Virginia.edu)

Wikipedia defines **cyber-harassment** as "a form of bullying or harassment using electronic means" [1]. Reading the definition, one would think of a platform such as Facebook, Instagram or any other social networking website, but unfortunately the issue is far too prevalent across all the online platforms [2]. And, nowadays people have accepted harassment as yet another feature of the internet. But, given how deeply digital life is intertwined with our physical life, the harassment can have a significant real-world impact, ranging from mental stress to fear of one's personal safety. Wikipedia counts openness as a key pillar [3], and even its content is entirely driven by people, so it realizes that ensuring a safe environment for its editors is important, as harassment disengages editors from the platform [2]. To prevent harassment, it has a strict no personal attacks policy, which specifies that offenders would get blocked [4]. Currently, to identify abuse Wikipedia relies on people to report harassers, whose actions are then evaluated by an administrator who performs the block if any blockable offense is detected.

Jorchi Ito described how **Singularitarians** are hopeful of a future, where machines would understand and solve problems by themselves [5]. As part of our capstone for Wikimedia foundation, although we are not creating a super-intelligent machine, we are creating an intelligent tool which would help reduce online harassment. Our goal is to predict users who are at risk of committing harassment using the system, improving both the turnaround time and harasser capture rate compared to the current human driven process. Given that the system has such a big mandate, the system's results would have huge implications, and I believe that the three key aspects to be taken into consideration are: **biases**, **opaqueness**, and **data privacy**.

Kate Crawford argues that data inherently contains hidden biases, either due to the creation of human design, or based on human interpretation among other reasons [6]. And, I believe **biases** could be present in our model as well. It becomes significant as the outcome of our model could directly lead to a user being blocked. And, with each incorrect prediction we lose a dedicated user and their trust, so the cost of every biased misprediction is very high. In our model, biases can broadly be due to misinterpretation of comments and inherent bias due to the lack of diverse perspective or a signal problem. Consider a case that certain users have a small representation in Wikipedia and their style of communication is a bit distinct, in such a case as the model wouldn't have encountered plenty of data about such a style earlier, it might be biased against them and might mark them as offensive more frequently. As a result, a large number of such users might get banned, and then, in turn, a model would not get adequate quantity of comments to understand their style of communication. The predicament seems very similar to the one captured by Cathy O'Neil when she argues about problems with predictive policing [7]. As in both scenarios we have a benign intention but our blind trust in the model doesn't help, as knowingly or unknowingly the model is biased, in turn affecting people's life. The problematic feedback loop in predictive policing stigmatizes the poor, eventually leading to a racial discrimination, whereas in our case it might lead to discrimination against a group of people with a similar communication style.

We can overcome this type of biases by taking proactive and reactive approaches. In a proactive approach, we could analyze if there are certain topics/pages where people have a higher propensity to get blocked and create awareness about the block policy so that people are extra vigilant about what they comment. In a reactive approach, we can give users an option to justify their action and then verify their claims, and if found valid we can notify the model about these incorrect predictions so that it doesn't misclassify them and corrects itself for future. Another inherent problem with predictive modeling is that, they are time bound. We might create a model based on current writing styles, and model would try to detect abuse based on prevalent slangs and other abusive words, but with time these words keep evolving. So, in future years, if the system is not updated it would not provide an unbiased prediction. A simple solution for this time bounded nature and data biasness would be to ensure that the system is routinely updated with recent data, so that the model keeps on learning the new words. An inherent problem with these solutions is that they ignore the gravity of impact each biased misprediction causes to a genuine user, as they lose their right to edit and comment on issues they care about.

Cathy O' Neil rightly mentions **opaqueness** as one of the weapons of math destruction [8], and to me it seems to be another important issue. As in our capstone, we are using a machine learning model to make the predictions, the inherent nature of the model is opaque so we would never be aware of what is driving the conclusions from the model. To understand the impact of the issue, consider a simple case of the model predicting whether to block a user or not by analyzing the number of special characters used, which in a way is a good metric as people often use them alongside abusive language. But, there could be a case where younger people or other groups of people who use special characters on a day to day basis might get incorrectly tagged by the model. So, if we knew how the model was making this incorrect prediction we could have corrected it instantly, but as the model is opaque we might not be able to fix it easily. John Danaher captures this issue about the output of machine learning algorithms being difficult to interpret, when he talks about algorithmic opacity [9]. He also illustrates the impact succinctly when he elaborates on the financial crisis, as to how the risk ratings were designed to be intimidating so that people could not comprehend them.

I believe opaqueness could be solved through multiple steps. Firstly, as pointed out by Yeshimabeit Milner [10], when we study the model in detail then one can identify the true impact of the model's prediction, so the data and model should be public. As, when a different set of people would look over the model, they would each bring in their perspective and as a whole help improve the model. Secondly, we should follow the FAIR principles [11], so that we ensure that every individual or organization has access to the shared information and there are no restrictions. Because, making data selectively available would be another form of discrimination. Thirdly, to create our model we can choose a machine learning technique which might be less accurate but is more inferable. But, the problem with choosing a less accurate model is that although we ensure fairness we lose out on the accuracy, so a lot of genuine users might get predicted inaccurately. Although, with the simpler model, we can understand the models functioning better so we could make the necessary suggestions to correct the biases.

While elaborating on the ten(limited rules), Mathew Zook et al mention that we need to guard against the reidentification of the data [12]. Although, Wikipedia puts emphasis on data openness, user **data privacy** seems like a key issue. As part of our capstone, we would be analyzing the entire history of user data in Wikipedia across the different platforms, so the data itself seems "human". As discussed above, to ensure there is no opaqueness and to align with Wikipedia's open platform policy, we would be ensuring that the model, results and predictions are available online. But, consider a case where we publish an anonymized list of users to be blocked based on our model, and to ensure openness we share their past history as well. From our end we have ensured user privacy by anonymizing the list of usernames. But, in today's world almost all anonymized data can be reidentified, and people who want to slander innocent users could use this as an opportunity to up their attack. So, there might be a case, where our inaccurate prediction leads to users' reputation getting tarnished, and this should be a big concern for us. The predicament seems similar to how Banksy's identity was revealed. In his case, researchers identified him using only the publicly available information and a model which was used to identify serial criminals [13]. And this is frightening, as even after knowing the magnitude of the public data available we underestimate how it can impact us. So, I believe that it is imperative we lay emphasis on user data privacy and protection in today's interconnected world. Given, how sensitive the matter is, I can't think of a clear solution as it seems there is a strict tradeoff between data openness and data privacy that needs to be taken into consideration. At the core of the issue, I believe information should be shared to the extent that it does not lead to reachability to the users. The European Union has taken note of data privacy and the General Data Protection Regulation protecting its citizens reflects that. I especially think that the "right to be forgotten" is a good step in that direction [14], as it gives the power of user data back into the hands of the user. To prevent misuse, EU ensures that only genuine user concerns which fit the legitimate ground rules are allowed to be forgotten.

Now, I believe that given how prevalent cyber harassment has become, and the magnitude of growth online platforms have seen, harassment would only be contained with the help of an automated system. And, while we create the system we need to ensure that it is unbiased, open, and respects data privacy. As, along with making Wikipedia remain a safe platform we need to ensure we respect user and their data.

References:

1. "Wikipedia : Cyberbullying". [Online]. Available: https://en.wikipedia.org/wiki/Cyberbullying. [Accessed: 11-Dec-2018].
2. "Online harassment. Pew Research Center, 2017". Available: http://www.pewinternet.org/2017/07/11/online-harassment-2017/. [Accessed: 11-Dec-2018]
3. "Wikipedia : Five pillars". [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Five_pillars. [Accessed: 11-Dec-2018].
4. "Wikipedia: No personal attacks". [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks. [Accessed: 11-Dec-2018].
5. "Resisting reduction". [Online]. Available: https://jods.mitpress.mit.edu/pub/resisting-reduction. [Accessed: 10-Dec-2018].
6. "the hidden biases in big data". [Online]. Available: https://hbr.org/2013/04/the-hidden-biases-in-big-data. [Accessed: 11-Dec-2018].
7. Cathy O'neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Penguin Books), 86-87.
8. Cathy O'neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Penguin Books), 31.
9. Danaher, John, The Threat of Algocracy: Reality, Resistance and Accommodation (Springer: Philosophy & Technology), 254-255.

10. “an open letter to facebook from the data for black lives movement”. [Online]. Available: https://medium.com/@YESHICAN/an-open-letter-to-facebook-from-the-data-for-black-lives-movement-81e693c6b46c. [Accessed: 11-Dec-2018].
11. “fairprinciples”. [Online]. Available:
https://www.force11.org/group/fairgroup/fairprinciples. [Accessed: 11-Dec-2018].
12. Zook, Matthew, Solon Barocas, Kate Crawford, Emily Keller, Seeta Peña Gangadharan,
Alyssa Goodman, Rachelle Hollander, Barbara A Koenig, Jacob Metcalf, and Arvind Narayanan, Ten simple rules for responsible big data research (PLoS computational biology 13 (3):e1005399).
13. “How scientists claim to have uncovered banksys identity”. [Online]. Available: https://www.marketwatch.com/story/how-scientists-claim-to-have-uncovered-banksys-identity-2016-03-05. [Accessed: 11-Dec-2018]
14. “Right to be forgotten”. [Online]. Available: https://gdpr-info.eu/issues/right-to-be-forgotten/. [Accessed: 11-Dec-2018]