An Ethical Reflection on User Privacy and Transparency of Algorithmic Blocking Systems in the Wikipedia Community

By - Charu Rawat (cr4zy@virginia.edu)

The English Wikipedia user community is one of the largest online user communities in the world. It has popularized the production of information and dissemination of encyclopedic knowledge around the world. With more than five million articles, its popularity as one of the five most visited websites in the world and its cultural significance as a free resource epitomizes the power of amateur online cooperation resting upon its principles of information freedom, user autonomy, and open-access policy [1].

As per the Wikimedia Community engagement insights it was observed that half of the respondents who experienced harassment on their platforms reported a decrease in their contribution and engagement levels. Cyber harassment has inhibited the open exchange of ideas and resources amongst Wikipedia's 34 million registered members spread across the world. To combat this problem, as part of my capstone project, I will be analyzing the Wikipedia user activity data publicly made available by the Wikimedia foundation and build a model using machine learning algorithms that can automatically detect abusive behavior and flag problematic users. The use of machine learning algorithms in decision-making processes are related to social issues with consequences. In this case, the consequence of committing the offense of harassment online is that the user gets "blocked" from the community and cannot contribute further to the platform. Because this consequence is grave and absolute, the entire process requires a thorough reflection in terms of ethics and, more broadly, of governance as well. I will be touching upon two issues fundamental not only to the ethics of any big data system nowadays but also which stand today as the core values upon which the Wikipedia user community has been built – transparency and privacy.

The focus on the functional development of such a system requires resources suitable for the ethical development and above all in line with the data it processes and the decisions it guides. One of the worst outcomes of such a process can be the introduction of distortion or flight of responsibility, from time to time referring to the cause of decisional errors to the algorithms instead of the decision makers.

There are potentially multiple issues that can be associated with an automated machine learning system that has been built to identify and predict abusive behavior or content based on historical data. In my case, my machine learning system will be trained upon "labeled abusive" data annotated by the human Wikimedia moderators. This raises the possibility of errors introduced by unintended bias or inconsistency on behalf of the moderators that flows into the data upon which my model could be trained. Kate Crawford rightly talks about the assumed objectivity that we expect out of bid data-

driven processes [2] and how hidden biases in both the collection and analysis stages present considerable risks as I described in this case. Another shortcoming of machine learning methods and algorithmic based decision making in detecting abusive content online its tendency to ignore message context but focus on content. The misinterpretation can lead to users getting blocked wrongly. Algorithms leveraged for such systems don't just analyze user activity and behavior but also study networks built by the user in the community through social interactions which could raise the risk of networked discrimination and affect already marginalized user groups in the community (Boyd et al., 2013) [3].

In the context of such issues, transparency becomes a fundamental prerequisite to avoid discrimination and solve the problem of information asymmetry, guaranteeing users the right to understand public decisions. It is also necessary to think about the policies chosen to determine the reference indices (benchmark policies) to avoid the effects of a larger dimension: just as an administrator can act in a non-transparent manner, pursuing not the common good but private interests, a non-transparent algorithm could carry out the same offenses even more broadly, producing not only injustices but also social discrimination.

User privacy and safeguarding the anonymity of users is one of Wikipedia's most cherished values. When designing a system of machine learning governance, some trade-offs are inevitable. For example, individual privacy considerations often must be balanced against the desire to achieve legitimate social ends. The extent to which specific values are embedded in systems reflects the priorities and preferences of the systems' designers. And the extent to which users accept and utilize these systems likewise reflects users' priorities. In my case, while the objective is to achieve a positive social goal, it does in no way outweighs the responsibility that I have when it comes to respecting user privacy, and that is to not use the data in unintended ways. While Wikipedia does offer its users the right to remain anonymous while indulging in editing activities, this doesn't completely guarantee the privacy of the strongest sorts that a user expects. Barocas and Nissembaum[4] make a strong case for why anonymization is only one way to bypass privacy issues but does not solve ethical issues related to it as anonymized data does not identify on the basis of name and address, but remains connected to a person. To illustrate the same in my project, analyzing decades worth of user activity including information such as IP that can approximate location, provides the ability to profile users in specific ways and reveal information pertinent to the user's true identity. For example, from the coordinates of articles that a user has edited, it is generally possible to determine the user's location even more precisely, or, time analyses of certain days of the year allow inferences to be drawn about a user's family status. It is probable, for example, that those who tend not to edit during the school holidays are students, parents or teachers. This raises the question that if we in any way are causing harm to the true identity of the user.

Crawford talks about leveraging social science methodologies to bring context-awareness to research to address serious signal problems in the data even though it makes the challenge of understanding big data more complex. To limit the bias in my data, and provide more context to an otherwise isolated language analysis process, I intend to use additional data available from Wikipedia that can aid my existing analysis by providing more context to user conversations that happen such as the original topic of conversation or the category a conversation falls under as opposed to just evaluating the conversation text. Other than the issues that arise out of such systems that can be potentially thought of in advance, often most problematic issues arise once the system or the process is in effect and is processing or ingesting unseen data in real time. In light of being transparent about the work done by me and my team with the Wikimedia Foundation and to adhere to Wikimedia's open access policy and core values, we intend to roll out our automated model to Wikipedia's user community for period of time during which it will be available for testing and users will have the ability to give feedback about it and point out any potential flaws that the model has. In addition to that, we are also now incorporating a feature into our model which when flags a problematic user will also give evidence along with it as to why that user should be blocked from the community. This will ensure that this does not remain a black box process but that there is hard evidence for action called. As designers of such a system, we have the ability to control the features engineered out of data that usually give more information about a user that what the data offers directly by making connections across various data points. To make sure that we are protecting user privacy and not leveraging data in unintended ways to reveal anonymity, as of now we plan to put constraints on the features developed during our model building process. We also intend to detail the methodology, thought the process and any assumptions involved behind the research into a document which will be available to the user community.

These measures don't offer complete solutions to all the potential misgivings of the data and the process. For example – certain users might disagree with the certain results offered by the model such as in a case where a user does indulge in some problematic behavior but not to the severity of being blocked. In such cases, there might be need for human intervention along with a review of the situation that warranted a user block. My sense is that such situations can often happen since at the end of the day since such a process is subjective, and subjectivity is difficult to incorporate into automated systems or models. And so there will always have to be an element of human judgment or interference involved with such process and that hundred percent dependency on such systems is dangerous.

Wikipedia prides itself in being a new age democratic society where people participate in making the decisions and have an equal right and say. Opacity is detrimental to the

Wikimedia Foundation's mission to promote the dissemination of free knowledge around the world and to its users. Transparency, as we all know, is key to any democratic system. Our hope is that the level of transparency that my team and I offer will help mitigate any unintended bias that could be caused by the model or the process and help improve what we've built. Overall, while the goal is to help protect the user community from harassment, in light of the ethical implications of the projects and measures taken to address it, I believe that the project can help the community gain a better understanding of itself, and that the transparency induced in the development and implementation of such an automated system will not only help increase the user community's understanding of it but also ours.

References

(1) Waskul, D., & Douglass, M. (1996). Considering the electronic participant. *The Information Society, 12*(2), 129–140.

(2) Kate Crawford (2013). *The hidden biases in big data*. https://hbr.org/2013/04/the-hidden-biases-in-big-data

(3) Paul Baker & Amanda Potts (2013) *'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms, Critical Discourse Studies*, 10:2, 187-204, DOI: 10.1080/17405904.2012.744320

(4) Barocas, S., & Nissenbaum, H. (2014). *Big Data's End Run around Anonymity and Consent*. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), Privacy, Big Data, and the Public Good: Frameworks for Engagement(pp. 44-75). Cambridge: Cambridge University Press. doi:10.1017/CBO9781107590205.004