An Ethics driven approach for implementing automated harassment blocks on the Wikipedia platform

Name: Arnab Sarkar
Email: as3uj@virginia.edu

Introduction

The advent of the Internet can be easily heralded as one of the key events which led to the Information age as it is colloquially known. Sharing of thoughts, ideas and opinions reached new heights and people of all ages, race, gender were able to engage in meaningful debates (though often using anonymized profiles), across chat rooms, social forums, and discussion boards. However, there was a darker aspect to this new medium – online abuse and harassment became rampant in certain communities. A 2017 survey by the Pew Research group reported that "Roughly four-in-ten Americans have personally experienced online harassment, and 62% consider it a major problem." [1] Wikipedia, since its inception in 2001, is no stranger to this phenomenon. In a recent Community insight survey in 2018, "68% of respondents reported having experienced harassment at some point in the past." In the same survey, "About 22% of Wikipedians reported avoiding Wikimedia projects for one to three days because they felt unsafe."[2] In order to combat this problem, Wikimedia has an organic, human-driven process in place, where people who have experienced harassment, report such occurrences on noticeboards, where these cases are acted upon by Wikipedia Administrators. However, since this a completely human-driven approach there have been cases of bias as well of neglect. It is the objective of our capstone project to develop a data-driven approach to automatically detect abusive users online and alert the human administrators of the Wikipedia Foundation of such abusive users. By making this an autonomous process, we intend to reduce the errors in judgment and also address the wide variety of issues, which cannot be addressed by a limited number of human administrators. However, there are a few ethical issues that need to be addressed before we go ahead with our implementation and these issues will be explored next in the paper.

Issue 1: Right to free speech and expression on online platforms

In most democratic countries, citizens have the right to free speech and expression, usually guaranteed by a charter or the constitution of the country in question. One common example cited is the US Constitution's 1st amendment. However, blocking/banning individuals for what they say on online platforms like Wikipedia seems at odds with the rights laid down by the country's constitution. This gives rise to questions – Can such a right which is guaranteed by governing institutions be overridden on the private online platforms? Should admins and moderators on such platforms like in our Wikipedia project case, be allowed to curtail the right to free speech on the Internet? The answer to these questions can be quite murky and most often than not comes down to the context of the speech and type of platform in question. Individuals frequently take advantage of the perceived anonymity of computer-

mediated communication, using this to engage in behavior that many of them would not consider in real life. Certain speech like bodily threats, incitement to lawless action, blackmail, obscenity, etc are not covered by these rights and can be grounds for prosecution. Lata Nott, an expert on the first amendment mentions that the first amendment only protects you from government punishing, censoring or oppressing your speech and does not apply to private organizations."So if, say, Twitter decides to ban you, you'd be a bit out of luck," Nott says."You can't make a First Amendment claim in court."[3] However, for a more public and open platform like Wikipedia, implementing bans/blocks is not as easy, as Wikipedia needs a collegial atmosphere needed to create a good encyclopedia.

A solution to this would be that Wikipedia should lay down firm rules in the form of community guidelines, borrowing from more private platforms like Facebook, Twitter, but also taking into account the unique open atmosphere at Wikipedia. The current principles laid down by Wikipedia are known as the "five pillars"; they are not enough to cover cases of harassment.[4] Even though Wikipedia has a "No personal attacks" policy and it's own Wiki-Etiquette[5], often human administrators in the absence of stricter guidelines have to arbitrate on their own. Most severe breaches of etiquettes can be easily judged, however without guidelines, it can be harder to arbitrate on more trivial offenses. For the purpose of implementing the solution proposed by our project, the Wikimedia Foundation would have to lay down strictly written guidelines, which can be adhered to by humans and machines alike. This would also make it clear to the public at large what speech would be considered offensive and grounds for banning by administrators and future autonomous methods, though there would still be a degree of bias involved – the paper would expand on this issue in the next sections.

Issue 2: Algorithmic Ideology - Who gets to decide the criterion for user blocks?

As mentioned earlier, for our implementation of autonomous user blocks, we will be using prior rules as well as new guidelines(if any) as the criterion for deciding whether a user's speech or online activity gets classified as block-worthy or not. These rules as well our study of previous cases of user blocks by the administrators will become the core of our algorithm for the automated block process. However algorithmic design and ideology are often influenced by certain biases, which are not conspicuous often. In our case, the Wikimedia Foundation will be giving us the data(of prior cases) and the rules for designing the algorithm and thus will to a certain extent influence its design. The criterion that they decide may be biased in many different ways. In our case we are building the new process for English Wikipedia, therefore most of the rule makers would likely be from the western world and most likely male. Hence, even though they think that the rules that they are envisioning for everyone, would most likely be more suitable for some "default" standards or the majority, and often this might be "men", "white", "straight", "rich" as assumed standards. As mentioned by Kate Crawford, this will be like the "signal problem": where data is assumed to reflect the social world, but there are significant gaps, with little or no signal coming from particular communities.[6] Another area of concern is Algorithm opacity where often it can be difficult to interpret the outcome of algorithms due to many factors involved. If the targeted people don't understand complex technologies, it can be easily be abused by the few "technological elite" who do understand them. As mentioned by Danaher, this is

essentially an "Algocracy" as a "particular kind of governance system, one which is organized and structured on the basis of computer-programmed algorithms."[7] Finally, since we are building models based on historical blocks data, we must take into account the fact that our models might become recidivist in nature, and have certain harmful feedback loops.

In order to address these concerns, Wikimedia foundation can take a few steps as the decision maker in this process. First, they can ensure that the team drafting the rules/guidelines for English Wikipedia harassment blocks has a diverse group of individuals from different races, ethnicities, genders and sexual orientation. This will not only fix the "signal problem" issue but also has the potential to add a lot of innovative ideas to the decision-making process. Algorithmic opacity can be handled to some extent by drawing a set of responses that the algorithm makes for different block cases/scenarios so that the user getting blocked knows the probable reason they were targeted for a block action. This is not a perfect solution, but generally more openness in response is the way to go forward here. The recidivism issue can only be handled if we are careful to address factors in the model which are chosen because of certain groups in the past that were associated with them. This can be something like geography(which is often proxy for race). We must do a careful analysis before selecting features for our model and make sure that model gets updated with more recent data points – only then will we avoid harmful, biased feedback loops.

Issue 3: Data Fundamentalism - Human data as the objective truth?

The data which we will be using for model training will be historical user blocks based on human interactions. However, such large-scale human data is bound to subjective – what one person found offensive may not be offensive to another. As Rebecca Lemov argues, "When trying to understand the ramifications of this big-data trajectory, I argue, it is necessary again to bear in mind that the data is not only generated about individuals but also made out of individuals. It is human data."[8] We are at this point essentially treating the data as a raw resource. However treating such data as a raw resource, in turn, leads us to think of it as the starting point of our analysis – and we tend to make the unnoticed assumption that the data is transparent, self-evident and that it is the fundamental truth itself. The question that should be asked here is – even though the human data may be biased, but due to its nature as a starting point resource, should such data be considered as the fundamental, objective truth and be used for training our model? This is the problem of Data Fundamentalism and it is best stated by Lisa Gitelman – "If we're not careful, in other words, our zeal for more and more data can become a faith in their neutrality and autonomy, their objectivity. "[9]

One way to tackle the Data Fundamentalism is simply being aware as designers of the model of the inherently subjective nature of the human data and to not treat any conclusions reached from big data models as what's really true. The results of a model should always be treated as just a simulation of what "might" be the truth. We should also never believe any correlations which we observed from the data and associate them as causal factors. Causality is only established through scientific, objective processes, independent of the collected data. Technological objectivity i.e. looking at how the data is produced, collected and the reason behind its collection, etc should all help us understand the underlying assumptions and biases.

There are some other measures which might help us understand the data source and the associated cognitive biases such as using big data features with qualitative methods and making use of ethnography while analyzing.

Conclusion

Through the implementation of our autonomous user blocks process, we hope to make Wikipedia a safer place to edit articles and interact with people who work in the same fields. Through our project, we would like to help maintain an atmosphere conducive to open knowledge sharing and collaboration on this platform. We are wary of the fact that even though we have good intentions, we should not use any means to reach those ends – in short, we should make sure that any Ethics of ultimate ends is balanced by Ethics of responsibility. Even though we would not be able to take into account all ethical issues, we feel that if we can account for the above-mentioned problem areas and be open to visualizing any new problems in our process, we would be able to positively impact a lot of people on Wikipedia.

References

[1] M. Duggan. “Online harassment. Pew Research Center, 2017”. Available: http://www.pewinternet.org/2017/07/11/online-harassment-2017/. [Accessed: 12/09/2018]

[2] “Wikimedia Community Engagement Insights, 2018 report.” Available: https://meta.wikimedia.org/wiki/Community_Engagement_Insights/2018_Report/Support_%26_Safety [Accessed: 12/09/2018]

[3] Lata Nott. “The First Amendment doesn't guarantee you the rights you think it does”. Available: https://www.cnn.com/2017/04/27/politics/first-amendment-explainer-trnd/index.html. [Accessed: 12/11/2018]

[4] “Wikipedia: Five Pillars”. Available: https://en.wikipedia.org/wiki/Wikipedia:Five_pillars. [Accessed: 12/11/2018]

[5] “Wikipedia: No Personal Attacks”. Available: https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks. [Accessed: 12/11/2018]

[6] Kate Crawford. “The Hidden Biases in Big Data”

[7] John Danaher. “The Threat of Algocracy: Reality, Resistance and Accommodation”

[8] Rebecca Lemov. “Big Data is People”

[9] Lisa Gitelman. “Raw Data is an Oxymoron”