6.804/6.864 HW2 TASK A: Comment Moderation

Social media platforms and other online forums are an increasingly common venue for discourse. But unfortunately, they have also created avenues for online harassment. A recent Pew Research Center survey found "41% of Americans have been personally subjected to harassing behavior online, and an even larger share (66%) has witnessed these behaviors directed at others."[1]

The effects of online harassment on the victim can be severe: from compromising their privacy to threatening their physical safety. And there are also significant negative consequences on the quality of these online platforms: both experiencing and simply witnessing harassment silences users and can ultimately drive them off online spaces. In its "milder" forms, toxic discourse can still foster a negative and hostile environment.

While employing moderators to read through and mark toxic comments is one potential mitigation, it's difficult to scale. Recent efforts have tried to utilize technology to help with comment moderation efforts — for example, by building machine learning models to identify abusive comments. These might be used in a variety of ways, from helping human moderators prioritize what to look at, to allowing readers to filter what comments they see.

These models are most likely trained on data consisting of past comments annotated by whether or not they contained toxic speech (or how much). In this part of the assignment, pretend you're on a research team at a new social media company that's collecting this data. You have a set of comments scraped from Wikipedia, and now you want to label them. You have a budget to hire annotators, but you need a labeling scheme and instructions to give them.

Here are some examples of what the comments look like:

- `Your comments on my discussion page are rude, arrogant, bullying and totally inappropriate. Napoleon complex is a stub and you might learn something about yourself by improving it, little boy.`
- `hello old and crazy man.....`
- `Prick. Gimme some time to flesh things out. Stop being such a prick.`
- `So when is someone going to warn YOU about your toxic attitude then? Tell me that. Tell me when someone's going to give you a block warning for the bullshit you've pulled on Wikipedia, like getting into arguments with people and issuing blocks to them when you don't like hearing the truth.`
- `Surely concerns re: fossil fuels/alternative energy sources, pollution and other environmental concerns should be included with the mention of animal rights.`
- `I have restored material that was removed and added a substantial reference within the text and also tried to add some perspective and also some copyediting. I would appreciate those of you who feel the`

[1] https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

information restored is unfounded taking a look at the references at the bottom of the page. With the opening of the KGB archives a flood of material has become available.

- Stop being such a goddamn prick. The article will be sorted out in time. Meanwhile spend some time away from wikipedia. And do normal things. for instance leave your parent's basement.
- You are full of shit sir. You are clearly one of those racists who lurk all over the internet looking to shut down viewpoints that differ from yours and expose truths covered in fantasy. There is NO reason that you should have blocked me or even said anything. What you should have done is to reply to what was written with intelligence. Since you did not, it is clear that you lack it. You are the prime reason why the internet community is finally realizing that this site is a joke and is covered by clowns with agendas. That is, people who want to put forth propaganda.
- USA is #1 and we made the facebook, deal with it. Oh wait, all you kids on facebook do instead is cry about things on the facebook. Make a stupid facebook group about it, why don't you.
- No! This is a GROUP EFFORT! Wikipedia is a collaborative COMMUNITY and there are no school essays here. The article needs to be more professional and adopt a better title besides the references. This is all that needs to be done, so get off your high horse and accomplish what you want to see done. If you have these goals, then put yourself to the test of solving this problem. That's what I do whenever something perturbs me. You're just looking for a fight about something you admittedly care nothing about. How about I come by your house and criticise your gardens? ``Why?`` You say. ``Because they are too ugly and I don't like the way they don't blend in with the neighbours' yards. So tacky, but I'm only passing by and I've never been down this road before.``

RES.TLL-008 Social and Ethical Responsibilities of Computing (SERC)
Fall 2021