Arthur Cheong, Andrew Yu, Marco Widodo - Group 10

Professor Mallory Nobles

DATASCI-154

19 August 2023

**Group 10 - Final Report**

With our research design, our objective is to minimize the racial discrimination faced by minority users on Twitter by exploring the following question: “How should Twitter balance Machine Learning with Human Input in their hate speech detection framework to minimize racial discrimination against minority users on their platform?” Our objective moves the needle on this social issue as it provides a better understanding of how Twitter—and other social media companies—can approach their hate detection systems to minimize discrimination on their platform and, by extension, foster a safer digital environment for their users.

Twitter is one of many social media platforms that incorporates a mix of human elements and machine-learning algorithms in their hate speech moderation. While this growing transition to algorithmic detection allows for greater speeds of moderation, it overlooks an underlying social problem; machine-learning algorithms trained on racially-biased data (Davidson, et al. 2019) become racially biased in their content moderation (Sap, et al. 2019). Social media companies have an obligation to their users to provide a platform safe from hateful content. Problematically, both ends of the spectrum—no automation and full automation—result in discrimination on the platform; more machine-learning involvement means faster processing capabilities but less accurate judgments, while more human input generally means more accurate judgments but slower processing capabilities, which translates to less discriminatory hate speech being filtered. There are a multitude of sources that have found evidence of discrimination both via hate speech on social media and through indirect silencing by algorithmic hate speech detection models; instead of tackling one or the other, our design seeks to confront the complex balancing act as a whole. We represent this problem using the following equation:

**find the minimum of:**

***RD = HI(a) + ML(b)***

In the above equation, *RD* quantifies the degree of racial discrimination experienced by minority groups; it is our objective to minimize this externality for increased social cohesion and online safety. As our main source of data, we plan to measure *RD—*and, by extension, the achievement of our objective—by issuing an anonymous survey on the Twitter platform that asks three questions:

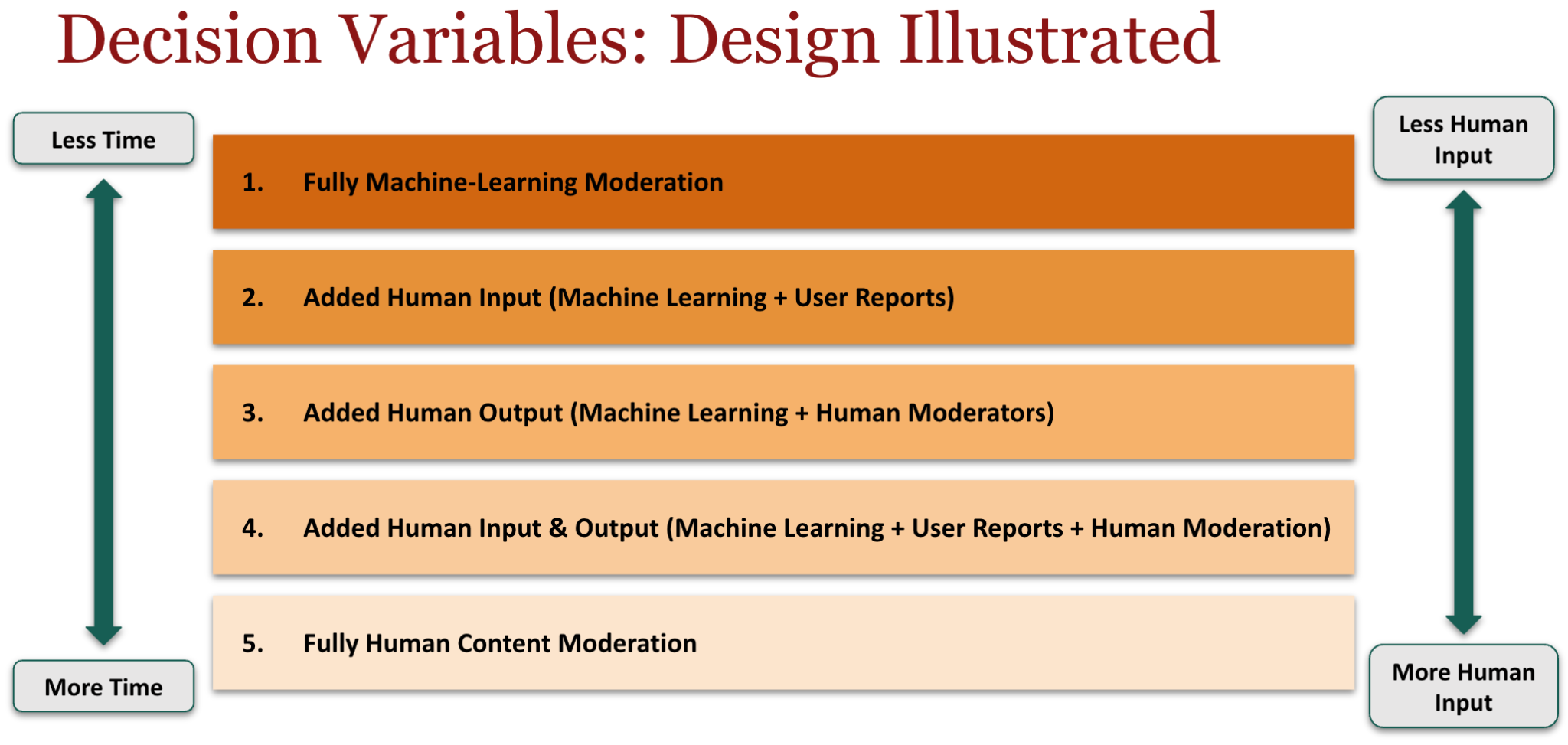
1. *“What Race/Ethnicity Do You Identify As?”*
2. *“On a scale of 1-10, to what extent do you feel you have been discriminated against by other users on Twitter?”*
3. *“On a scale of 1-10, to what extent do you feel you have been discriminated against by Twitter’s content moderation policies?”*

We can average the scores of Q2 and Q3 and group responses by their answers to Q1 to provide a rough measure of how much racial discrimination is experienced by each minority group. These average scores would necessarily fall upon a synthetic qualitative scale from 1-10. That said, our measurement of *RD* is constrained cross-spatially; we can only test on the entirety of Twitter’s user base at once since testing different sub-groups may yield significantly different results.

Our decision variables, *HI* and ML, represent the degree of Human Input and degree of Machine-Learning involved in the content moderation system, respectively. Our optimization model recognizes that a centralized decision maker cannot control individuals’ actions or choices; realistically, Twitter cannot directly reduce how much discriminatory hate speech is posted on their platform. But, optimizing the balance of these decision variables will meaningfully impact our objective as it may significantly reduce how much hateful content is visible, thereby minimizing its impact on users.

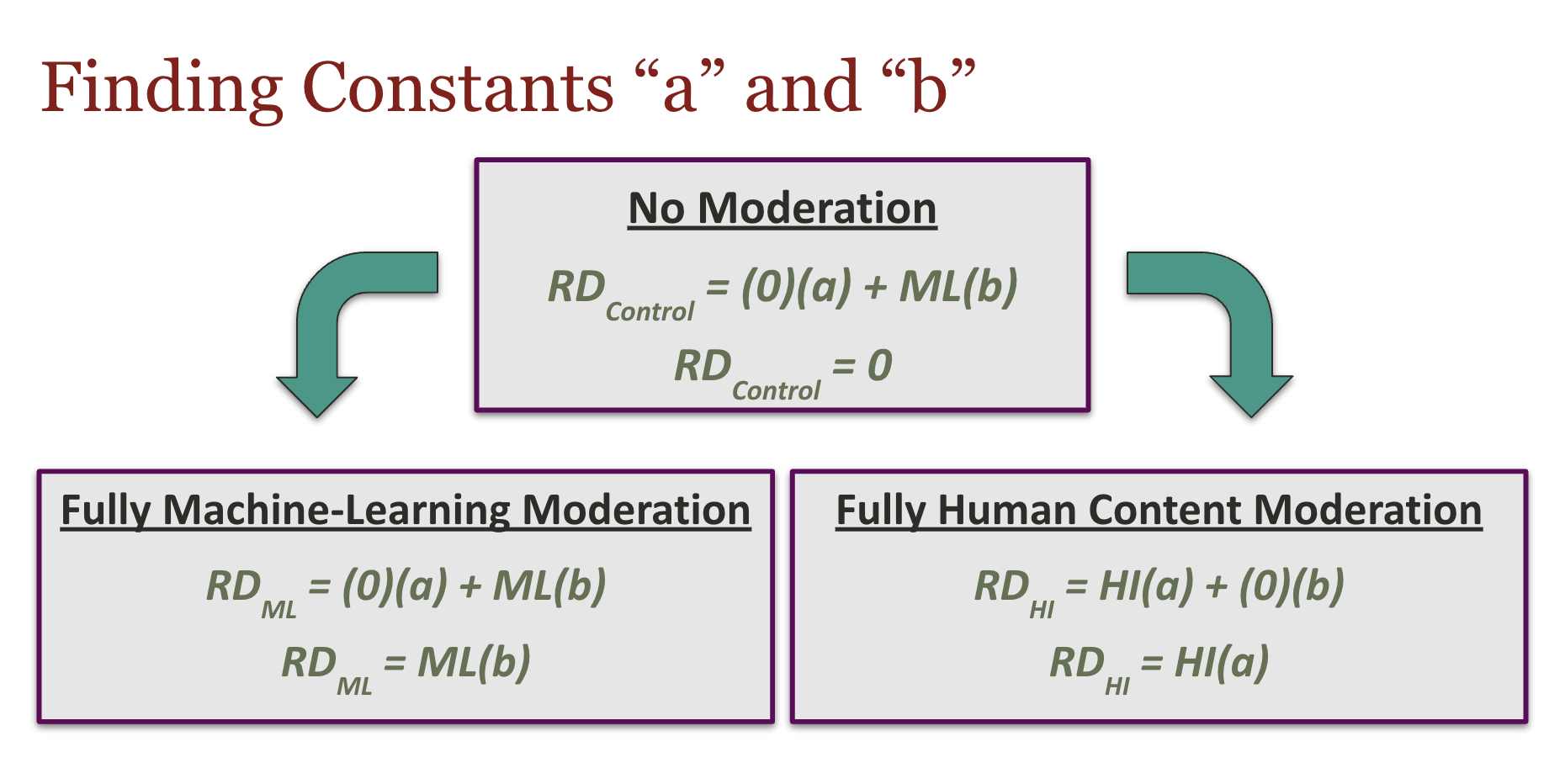
That said, top-down obstacles hinder reforms towards optimization as Twitter’s incumbent CEO—Elon Musk—has severely limited the power of Twitter’s human moderators, relying on algorithms instead. Through involving affected stakeholders in these anonymous surveys, they are prompted to reflect on discrimination on Twitter’s platform. In doing so, we raise public awareness about the prevalence and sources of digital discrimination, thereby increasing public pressure on Musk and incentivizing platform-wide moderation reforms.

When manipulating these variables for optimization, measuring the balance of HI and ML involvement is difficult to quantify in neat numbers because they function as parts of the larger Twitter ecosystem; we cannot choose some tweets to be sorted by human content moderators (HCMs) while the rest get sorted by machine learning (e.g. 30% human, 70% machine learning). Thus, we have created a spectrum ranging from more machine-learning involvement (1) to more human input (5):



As for how granular our measurements would be, we plan to rebalance and administer surveys once every month so as to allow the content on users’ feeds to reflect the output of the hate detection system being tested at that time. In changing our decision variables, one main limitation with our data may be a cross-temporal limitation as some months may yield skewed results; for instance, users may be more attentive of racial discrimination on the platform during Black History Month than during other months. As such, the research design will have to cycle over the five configurations multiple times (with each full cycle taking around 5 months) and average the measured *RD* scores.

Constants *a* and *b* represent how much human input and machine learning contribute to racial discrimination, respectively. These can be measured by comparing the RD value when no moderation is present to the RD value when moderation is fully human content moderation (*a*) or fully machine-learning (*b*).



In manipulating our decision variables, we face two primary constraints. First, the system has to meet a minimum efficiency rate, wherein the average time it takes for tweets to be uploaded should be more than or equal to the average collective time necessary for a tweet to be processed by the system:

**TUP  ≥ (THI + TML)**

What this entails is replacing human input with machine learning involvement is only feasible to the extent that the hate speech detection system as a whole can keep up with the average rate of tweets being posted. This is because social media companies like Twitter need to find hate speech detection systems that can be implemented relatively sustainably; Twitter can’t just shut down their platform every once in a while to work through a backlog of tweets in search of hate speech.

Second, the number of HCMs must be less than or equal to the total number of employees at Twitter, and greater than 0 because we cannot have a negative amount of human content moderators:

**0 ≤ PHCM ≤ PE**

To alleviate this constraint and related concerns about manpower, we aim to incorporate user reports to sieve out the tweets that are more likely to contain racial discrimination. Additionally, as part of our information to action design, we can implement short, mandatory modules for both new and current Twitter users to teach them how to report specific posts or accounts. Through this collaborative data collection, the human input portion of this balance will be supplemented by Twitter users rather than being predominantly run by HCMs. As Twitter users become effective at correctly reporting instances of hate speech on the platform, the overall efficiency of the human input portion becomes less limited by the availability of actual HCMs.

Potential data quality limitations for the surveys are two-fold. First, this design measures discrimination across distinct racial differences. Surveys used in iterations of this model would require adjustments to investigate Twitter discrimination experienced by those of mixed race or discrimination across different religions, nationalities, ages, etc. Second, it is also important to acknowledge the inherent limitations of surveys. These include the absence of repercussions for the provided responses and the constraint of limited attention. As a result, users might opt for satisfactory rather than well-reflected responses, and there is also the potential for responses to be influenced by a desire to appear socially desirable.

Another source of data we will be indirectly utilizing is the constant stream of tweets, starting from the date of implementation of this design. There are two main ethical considerations with the collection of this data. The first would be the failure to exercise caution in the creation of HCM teams, which may yield skewed and biased results, since humans inherently make decisions from their perspective and with the information they have; we would need to take care to ensure diverse backgrounds and equal representation of as many groups as possible (ethnically, religiously, socioeconomically, etc) from the onset. The other would be exposing human moderators to unfiltered content that may be graphic or triggering, potentially causing mental trauma; we would require explicit warning and consent for prospective HCMs, as well as urge Twitter to provide access to mental health care or therapy both during and after the research.

Bibliography

Amnesty International. (2022, May 12). Toxic twitter - the reporting process. https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-4-4/

Davidson, T., Bhattacharya, D., &amp; Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. Proceedings of the Third Workshop on Abusive Language Online. https://doi.org/10.18653/v1/w19-3504

Ghaffary, S. (2019, August 15). The algorithms that detect hate speech online are biased against Black people. Vox. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

Kovács, G., Alonso, P., &amp; Saini, R. (2021). Challenges of hate speech detection in social media. SN Computer Science, 2(2). https://doi.org/10.1007/s42979-021-00457-3

Leffer, L. (2022, November 1). Twitter’s content moderation team reportedly unable to work amid musk takeover. Gizmodo. https://gizmodo.com/twitter-content-moderation-musk-midterms-1849726968

Sap, M., Card, D., Gabriel, S., Choi, Y., &amp; Smith, N. A. (2019). The risk of racial bias in hate speech detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1163

Twitter. (2021, August 10). Combatting online racist abuse: An update following the euros. https://blog.twitter.com/en\_gb/topics/company/2020/combatting-online-racist-abuse-an-update-following-the-euros

Vanian, J. (2022, November 2). Twitter reportedly limits employee access to content moderation tools as midterm election nears. CNBC. https://www.cnbc.com/2022/11/01/twitter-reportedly-limits-employee-access-to-content-moderation-tools-.html#:~:text=Twitter%20is%20still%20using%20automated,%2Dprofile%20violations%2C%20Bloomberg%20said.