An Urgency for Inclusivity: Redesigning Datasets for Improved Representation of LGBTQ+ Identity Terms in Artificial Intelligence (A.I.)

Lani Wang
HSS4 - The Modern Context: Queer Theory and Politics
Professor Barnick
August 14, 2023

The Common Crawl Dataset stands as a vital resource in the realm of A.I. model training, enabling advancements in various fields through its vast collection of web page text, metadata extracts, and data extracts.[1] As A.I. continues to evolve, the representation of identity terms within the datasets it is trained on becomes increasingly significant, shaping the way these models perceive and interact with the world. However, despite the Common Crawl's prominence, handling LGBTQ+ identity terms in A.I. model training using cleaned versions of it has been deemed inadequate, giving rise to concerns regarding inclusivity and accuracy.[2] This inadequacy underscores the urgent need for community-led efforts to redesign A.I., prioritizing improved representation that diverges from corporate-driven projects. In this context, exploring the complexities surrounding identity terms and their representation in cleaned datasets becomes paramount to fostering a more inclusive and equitable A.I. landscape.

Data cleaning is particularly apparent in Google's C4 dataset, a filtered version of the Common Crawl.[3] The acquisition of C4 was based on a set list of heuristics, including the statement, "We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words"".[4] While this filtering approach aims to remove offensive or inappropriate content, it inadvertently eliminates pages that contain the accurate usage of non-normative identity terms. For instance, the inclusion of words like "sex"[5] and "sexuality"[6] on this list results in the removal of content that does not necessarily imply explicit or inappropriate

---

[1] Common Crawl, "The Data," Common Crawl Foundation, accessed August 9, 2023.
[2] Martin Anderson, "Minority Voices 'filtered' out of Google Natural Language Processing Models," Unite.AI, December 10, 2022.
[3] Google, "Papers with Code - C4 Dataset," C4 Dataset | Papers With Code, accessed August 9, 2023.
[4] Google, 6.
[5] Jacob Emerick, "LDNOOBW / List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words," GitHub, accessed August 10, 2023, 307.
[6] Emerick, 313.

material. Pages that discuss diverse perspectives and challenge traditional assumptions about sex or sexuality that mention either of these terms are eliminated, thereby erasing examples of genuine usages of non-normative sex and sexuality identity terms. By also excluding listed words associated with sexual anatomy and sexual health such as "vagina"[7] and "Viagra",[8] pages that contain critical medical knowledge surrounding these subjects are withdrawn. This exclusion leads to inaccuracies or misinformation in A.I.-powered diagnoses, treatment recommendations, health information, and other forms of digitally based healthcare. Individuals who are facing unique or uncommon health issues based on being of non-normative or non-biological sex are much more susceptible to these disparities in virtual aid. Consequently, A.I. models trained on such cleaned datasets are significantly limited in their ability to understand and represent the nuanced experiences of individuals identifying with these terms, weakening their overall ability to support these individuals through general automated applications such as algorithmically generated search engines or smart speaker voice assistance.

In practice, large language A.I. models built with the C4 dataset, such as Google's LaMDA, have demonstrated a lack of such support. Timnit Gebru, former head of Google's ethical A.I. team, and several of her colleagues describe problems with similarly built models and how they pose inherent risks, an inability to fully understand the concepts from their training data as well as the possibility to generate false claims.[9] Gebru and her team emphasize that with filtered data, the voices of people with a ruling or dominant perspective are more likely to be intact after filtration, and in the case of English text samples, "white supremacist and

---

[7] Emerick, 379.
[8] Emerick, 381.
[9] Emily M. Bender et al., "On the Dangers of Stochastic Parrots," Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 610.

misogynistic, ageist, etc. views are overrepresented in the training data".[10] Sexual minorities are constructively silenced in this majority-takes-all procedure, amplifying the already entrenched issues of silencing the unrepresented. This silencing also reaches these models' outputs, as they become more likely to respond with less mention and reference to LGBTQ+ voices and individuals as well as possess a disregard and avoidance for user input if it contains any content that it associates with LGBTQ+ individuals and experiences.

In the same vein, even if a model's data cannot be cleaned, its filtering method may be inherently cissexist in nature. As pointed out by Os Keyes, a Ph.D. Candidate at the University of Washington's Department of Human-Centered Design & Engineering, Automated Gender Recognition (A.G.R.), a sect of facial recognition that aims to identify an individual's gender based on photographs or videos, consistently produces trans-exclusive categorization from project to project.[11] A.G.R. algorithms are typically trained on datasets that uphold conventional binary constructs of gender, thereby reinforcing the notion that gender is a simple dichotomy between the male and female sexes. Furthermore, these algorithms are designed to categorize images or videos based on the perceived genders of their creators. Keyes's research reveals that gender is treated as a binary concept in 94.8% of papers, often assuming that gender can be confined to just two categories.[12] This perspective overlooks the intricate tapestry of gender identities, effectively erasing non-binary, genderqueer, and transgender individuals from consideration. Despite being employed in applications such as CCTV surveillance, security, and biometric systems, A.G.R.'s approach of assigning gender to only one of two classes inevitably falls short for those whose identities exist beyond the binary spectrum. This issue is also found in

[10] Bender, 613.
[11] Os Keyes, "The Misgendering Machines," Proceedings of the ACM on Human-Computer Interaction 2, no. CSCW (November 2018), 88:1.
[12] Keyes, 88:7.

the machine translation of non-gendered pronouns across languages. As pointed out by Sourojit Ghosh and Aylin Caliskan, a third-year Ph.D. Candidate and an assistant professor at the University of Washington respectively, the popular chatbot ChatGPT fails to translate the English gender-neutral pronoun "they" into the gender-neutral pronouns of other languages.[13] This is particularly reflected in languages that are understudied in the translation space such as Bengali. Although it is the seventh most spoken language in the world, machine translations of gender-neutral pronouns from Bengali into English, the language with the highest number of resources in A.I. training data, are unwittingly inferred a gender based on the user's text input.[14] Events, texts, or stories that contain gender-neutral pronouns in languages that appear in lower quantities within A.I. training data are at high risk of getting mistranslated, misinterpreted, and misrepresented by these algorithms, and by extension are at a high risk of losing the accurate meanings of their content.

A prime example of this misinterpretation and misrepresentation can be observed through the outputs of Google's pre-trained A.I. model google/t5-v1_1-small, which is a filtered model trained only on the C4 dataset.[15] When I prompted the textbot I built on this model with text containing LGBTQ+ language such as "tell me about queer history and context", the model would usually refuse to elaborate and provide short-winded responses like "No!!." and "?? OK I say. Definitely not!" In comparison, the outputs of Google's pre-trained A.I. model google/flan-t5-small, an unfiltered model trained on an uncleaned series of unpublished books, question explanations, and K-8 mathematics word problems, exhibit a greater degree of

[13] Sourojit Ghosh and Aylin Caliskan, "ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages," Arxiv, August 8, 2023, 1.
[14] Ghosh and Caliskan, 2.
[15] Colin Raffel et al., "Google/T5-V1_1-Small · Hugging Face," google/t5-v1_1-small · Hugging Face, February 12, 2020.

explicitness and elaboration.[16] When I prompted the textbot I built on this model with the same text as t5-v1_1-small, the model would output lengthier responses that were typically in some harmful or irrelevant manner towards LGBTQ+ individuals such as “one queer woman was a womanizer who slept alone with a group of people.” and “killed in a car accident in the wake of the events on the street,” In comparing the two outputs, it becomes apparent that flan-t5-small's unfiltered nature allows it to generate more comprehensive responses, even if they are more explicit. However, it is essential to note that the learning mechanisms of both textbots, t5-v1_1-small, and flan-t5-small, are inherently binary. Both bots are initially taught to learn by responding to a user’s text input with five different answers. The user then picks the answer they are biased towards and provides more text input. The bots then treat that answer as their “best” answer and then branch five more answers based on their “best” in response to the new user input. Selecting the "best" one aligns with a singular perspective, undermining the principles of queer theory. This binary nature limits the exploration of multiple viewpoints and stifles the recognition of alternative ways of thinking for the textbots. Although flan-t5-small appears to offer more depth in its responses, both models fall devastatingly short of capturing accurate, just, or neutral responses to text that contains LGBTQ+ identity language. This reinforces the urgent need for more inclusive training data and methodologies that genuinely reflect the complexity of LGBTQ+ identities.

In response to the limitations and biases inherent in these corporate, capitalized A.I. projects, community-led efforts to redesign A.I. have emerged. These community-driven initiatives prioritize diversity and inclusivity over a model or its training data and strive to develop datasets, models, and tools that truly represent the diversity of human experiences. As

[16] Alexandre Lacoste et al., “Google/Flan-T5-Small · Hugging Face,” google/flan-t5-small · Hugging Face, October 21, 2019.

noted by Gebru and her team, “size doesn’t guarantee diversity”.[17] While corporate projects may have the financial stability to conduct large-scale data collection and A.I. model processing power, they tend to overlook areas of the internet that have fewer links or are harder to reach. These areas, such as anti-ageist blogs, contain substantial discussion about the experiences of minority individuals and are impactful resources for capturing high-quality actual data on these groups. Community-led efforts are typically composed of people who know of these hidden or alternative pages and content and are less likely to be filtered out or shut down by a corporation’s mission to stay neutral to appeal to its market. These efforts are much more capable to carry out the collection of diverse, niche, quality data that can contain vastly detailed, insightful, and real cases to existing technological tools as well as find faults in current algorithms that have led to previously mentioned offensive and vague responses and output in corporate models in the first place.

A notable grassroots organization, Queer in AI, has been actively raising awareness around these queer issues in A.I. and machine learning. Their departure from a collaboration with Google was motivated by Google's failure to address “the harm they’ve caused by undermining both inclusion and critical research”.[18] This exemplifies the driving difference between community-led AI projects and corporate, capitalized ones in several crucial ways. The statement calls attention to Queer in AI's commitment to inclusion and representation of marginalized communities, actively seeking to address the harm caused by biases and exclusion in A.I. models.[19] It reflects the community-led initiative’s dedication to rigorous and thoughtful research, advocating for transparency, accountability, and responsible use of A.I. On one hand,

[17] Bender, 613.
[18] Black in AI, Queer in AI, and Widening NLP, “Statement to Google,” Queer in AI, May 10, 2021.
[19] Queer In AI, “Mission,” Queer in AI, accessed August 10, 2023.

sex and sexuality are often viewed as personal attributes, qualities, and values that differ from person to person. On the other, research is seen as a collaborative effort, one that requires the work of multiple others, organizations, funds, and legislation to arrive at substantial results. The nature of how typical A.I. research is conducted is something that cannot reflect the complexity and personal attributes of LGBTQ+ individuals and other minority groups. The particular needs of each individual are strongly considered in the prioritization of social impact that community-led projects have, focusing on using A.I. for the greater good and benefiting marginalized communities. Additionally, other community-led initiatives like PartnershipOnAI focus on real-time impact, creating tools that address problems experienced in the world due to A.I. Their projects such as the A.I. Incident Database serve as a central repository of such issues, promoting transparency and accountability.[20] This crowdsourced database keeps track of the collective history of harm or near harm dealt by A.I. that someone has found to be a threat.[21] The open access and simple user interface of this project allow for a clear understanding of the need for awareness and community involvement, even if one may not be familiar with A.I. It also provides A.I. practitioners examples of A.I. that led to a faulty outcome to avoid the chance of a similar mistake in the future. Furthermore, PartnershipOnAI also conducts and publishes public research to encourage the development of inclusive, human-centered A.I. They released a paper that details four guiding principles for ethical engagement in the production of A.I. alongside three recommendations aligned with those principles for building inclusive A.I.[22] While the statements listed can be associated with the workings of a community-led effort, they also act as active advice for corporate A.I. projects as well. PartnershipOnAI may not be as well known or

[20] PartnershipOnAi, "Ai Incidents Database," Partnership on AI, July 29, 2022.
[21] PartnershipOnAI, "Welcome to the Artificial Intelligence Incident Database," AI Incident Database RSS, accessed August 10, 2023.
[22] Tina Park, "Making AI Inclusive: 4 Guiding Principles for Ethical Engagement," Partnership on AI, July 20, 2022.

as highly backed as larger, corporate organizations like Google, but they have found aspects of current A.I. applications that would otherwise go unnoticed or disregarded by corporations. Corporate-backed A.I. projects prioritize profit-making objectives over social impact and do not always address potential harm or prioritize inclusivity and critical research. This stark difference in values and priorities emphasizes the crucial role that community-led A.I. projects play in fostering inclusive, ethical, and socially responsible contributions to A.I. as we see it today.

A critical comparison between community-led initiatives and corporate A.I. projects lies in their data collection methods. For instance, Google's Project Respect merely asks for people's identity terms without considering the real-life context behind those terms.[23] This approach results in biased and misrepresented algorithms, as it solely focuses on positive identity statements, failing to accurately represent all of their uses in real life. These statements do not come from discussion or actual context, providing models with very meager material to learn and work with for quality responses. Someone who is struggling with issues associated with certain identity terms will likely find it hard to work with A.I. trained on data from Project Respect without additional datasets that use those terms in a life-like setting. In contrast, initiatives like Queer in AI and PartnershipOnAI place accuracy and genuineness at the core of their missions, understanding that without it, true representation in A.I. is unattainable. By prioritizing quality data, community-led projects foster more accurate and equitable A.I. that reflects the complexity of human sex, sexuality, identification, and experiences.

Despite the benefits of community-driven A.I. projects, challenges also exist. Scaling up community-led efforts to compete with corporate-backed projects is difficult due to constraints

[23] Google, “Project Respect,” Google, accessed August 10, 2023.

such as funding, and access to data.[24] Additionally, ensuring representation and inclusivity within community-led projects requires ongoing vigilance and sensitivity to diverse perspectives. However, even with those shortcomings, community-led efforts to redesign A.I. stands as a powerful alternative to corporate, capitalized A.I. projects, emphasizing inclusivity and accurate representation of diverse identities. Organizations like Queer in AI and PartnershipOnAI exemplify the commitment to prioritize inclusivity, promoting transparency, and ethical engagement. While challenges remain, these community-driven initiatives pave the way for a more equitable and inclusive A.I. ecosystem.

The inadequate handling of LGBTQ+ identity terms in A.I. model training using cleaned datasets undermines inclusivity and accuracy, necessitating urgent community-led efforts in A.I. redesign. Overlooking LGBTQ+ and other minority group representation in A.I. development perpetuates biases, and hinders inclusivity, and marginalized communities. To rectify this, researchers, developers, and local communities must collaborate to prioritize reinforced diversity and inclusivity in datasets used in A.I. development. By actively involving marginalized groups in A.I. model design, data collection, training, and evaluation, we can ensure a broader representation of perspectives. Transparency and accountability are also crucial in openly addressing biases and limitations to garner trust. Fostering an inclusive A.I. ecosystem requires collective action, steering away from solely corporate-driven endeavors, and promoting A.I. technologies that empower and serve all users equitably.

[24] Anne-Kathrin Schwab and Rebeca Roysen, "Ecovillages and Other Community-Led Initiatives as Experiences of Climate Action," Nature News, June 12, 2022.

Bibliography

Primary Sources

Black in AI, Queer in AI, and Widening NLP. “Statement to Google.” Queer in AI, May 10, 2021. https://www.queerinai.com/statement-to-google.

Common Crawl, "The Data," Common Crawl Foundation, accessed August 9, 2023. https://commoncrawl.org/the-data/#:~:text=The%20Common%20Crawl%20corpus%20contains,cloud%20platforms%20across%20the%20world.

Emerick, Jacob. “LDNOOBW / List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.” GitHub. Accessed August 10, 2023: 307, 313, 379, 381. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en.

Google, “Papers with Code - C4 Dataset.” C4 Dataset | Papers With Code. Accessed August 9, 2023: 6. https://paperswithcode.com/dataset/c4.

Google. “Project Respect.” Google. Accessed August 10, 2023. https://projectrespect.withgoogle.com/.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. “Google/Flan-T5-Small · Hugging Face.” google/flan-t5-small · Hugging Face, October 21, 2019. https://huggingface.co/google/flan-t5-small#bias-risks-and-limitations.

PartnershipOnAI. “Welcome to the Artificial Intelligence Incident Database.” AI Incident Database RSS. Accessed August 10, 2023. https://incidentdatabase.ai/.

Queer In AI. “Mission.” Queer in AI. Accessed August 10, 2023. https://www.queerinai.com/mission.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. “Google/T5-V1_1-Small · Hugging Face.” google/t5-v1_1-small · Hugging Face, February 12, 2020. https://huggingface.co/google/t5-v1_1-small.

Secondary Sources

Anderson, Martin. “Minority Voices ‘filtered’ out of Google Natural Language Processing Models.” Unite.AI, December 10, 2022. https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-

models/.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots.” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 1, 2021: 610, 613. https://doi.org/10.1145/3442188.3445922.

Ghosh, Sourojit, and Aylin Caliskan. “ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages.” Arxiv, August 8, 2023: 1, 2. https://arxiv.org/ftp/arxiv/papers/2305/2305.10510.pdf.

Keyes, Os. “The Misgendering Machines.” *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (November 2018): 1–22. https://doi.org/10.1145/3274357.

Park, Tina. “Making AI Inclusive: 4 Guiding Principles for Ethical Engagement.” Partnership on AI, July 20, 2022. https://partnershiponai.org/paper/making-ai-inclusive-4-guiding-principles-for-ethical-engagement/.

PartnershipOnAI. “Ai Incidents Database.” Partnership on AI, July 29, 2022. https://partnershiponai.org/workstream/ai-incidents-database/.

Schwab, Anne-Kathrin, and Rebeca Roysen. “Ecovillages and Other Community-Led Initiatives as Experiences of Climate Action.” Nature News, June 12, 2022. https://www.nature.com/articles/s44168-022-00012-7.