Wiki Workshop (11th edition) – June 20, 2024

Supporting Community-Driven Data Curation for AI Evaluation on Wikipedia through Wikibench

Tzu-Sheng Kuo
Carnegie Mellon University

Aaron Halfaker
Microsoft

Zirui Cheng
Tsinghua University

Jiwoo Kim
Columbia University

Meng-Hsin Wu
Carnegie Mellon University

Tongshuang Wu
Carnegie Mellon University

Kenneth Holstein
Carnegie Mellon University

Haiyi Zhu
Carnegie Mellon University

Abstract

AI tools are increasingly deployed in community contexts. However, datasets used to evaluate AI are typically created by developers and annotators outside a given community, which can yield misleading conclusions about AI performance. How might we empower communities to drive the intentional design and curation of evaluation datasets for AI that impacts them? We investigate this question on Wikipedia, an online community with multiple AI-based content moderation tools deployed. We introduce Wikibench, a system that enables communities to collaboratively curate AI evaluation datasets, while navigating ambiguities and differences in perspective through discussion. A field study on Wikipedia shows that datasets curated using Wikibench can effectively capture community consensus, disagreement, and uncertainty. Furthermore, study participants used Wikibench to shape the overall data curation process, including refining label definitions, determining data inclusion criteria, and authoring data statements. Based on our findings, we propose future directions for systems that support community-driven data curation.

Keywords: artificial intelligence, community-driven AI, data curation, AI evaluation, English Wikipedia

Motivation

AI tools are increasingly deployed in *community contexts*. For example, AI-based content moderation tools have been deployed in online communities such as Wikipedia and Reddit. AI-based decision-making tools have also been adopted by local governments to prioritize public services, such as allocating local housing resources. However, the datasets used to evaluate AI performance are typically designed, curated, and labeled by developers and data annotators outside of a given community, which can lead to misleading conclusions about AI systems' "fit for use". In turn, the deployment of poorly-fit AI tools can yield compromised user experiences or even cause harm to vulnerable populations. For example, research shows that crowdsourced datasets systematically label innocuous phrases in African American English (AAE) dialects as toxic. As a consequence, if such datasets were used to prospectively evaluate content moderation tools' fit for use in a community that uses AAE, they would *underestimate* the tools' false positive rates, compared with what the community would experience in deployment.

Given that what constitutes "good performance" on tasks such as content moderation can be highly community-specific, recent work has argued that HCI and machine learning research should explore more *community-driven* approaches to AI dataset development. For instance, in a position paper, (Jo and Gebru, 2020) propose that AI should draw lessons from archive and library studies, where archives are often directly contributed and curated by the communities they are meant to represent, instead of by community-outsiders. These community archives, such as the Feminist Archive and the Working Class Movement Library, are motivated by the need to represent the voices of non-elites and the marginalized. The authors argue that these traditions should inspire new approaches to AI *data curation* that allow communities greater voice in specifying their collective desires for AI performance.

In the context of AI evaluation, *data curation* refers to the process of designing the "ground truth" against which AI models' performance will be evaluated. This involves an intentional process of selecting *which data points* should be included in a dataset and, in the case of labeled datasets, deciding *how each data point should be labeled*. For example, when developing an AI dataset for content moderation tools on Wikipedia, a "data point" could be an edit to an article, and its "label" could be a judgment of whether the edit should be considered "damaging" to the article or not. The intentional curation of

AI evaluation datasets stands in stark contrast with what Jo and Gebru term “laissez faire” approaches to dataset development, which indiscriminately take data in masses by crawling trace data on the web. On their own, such datasets simply capture how people *have behaved* in the past. However, they often fail to capture communities' normative beliefs about how decisions *should be made*, for evaluation purposes.

Related Work

Realizing the vision of *community-driven data curation* of AI datasets in practice poses numerous open challenges. For example, while a community may share broad norms and values, individual community members may disagree about how specific data points should be labeled (e.g., whether a given post should be considered “toxic”). In some cases, these disagreements may represent substantive differences in perspective, while in other cases, a brief discussion between individuals could reveal that they actually agree more than they disagree. Current approaches to account for annotator disagreement in crowdsourced datasets tend to handle disagreements post-hoc (after data have already been labeled), either by resorting to the majority vote or by attempting to model individual subjectivity for re-weighted voting. However, when it comes to deciding how important community decisions should be made, it is crucial that community members have opportunities to collectively build meaning and understand each other's perspectives. In contrast to prior methods, this calls for more collaborative and deliberative approaches that allow community members agency in navigating disagreements, via processes that are perceived to be fair by community members. Furthermore, beyond selecting and labeling individual data points, it is critical to provide communities with the agency to shape higher-level decisions, such as crafting label definitions and determining data inclusion criteria. Finally, given that community members will generally have limited time and attention to contribute to the curation of AI datasets, it is important to support them in prioritizing their efforts. To the best of our knowledge, despite recent calls-to-action from the research community, there are no existing tools aimed at addressing these challenges to support the intentional, community-driven curation of AI datasets in practice.

Our Method

We identify and address these challenges in the context of Wikipedia, an online community where multiple AI-based content moderation tools have been deployed, but where community members currently have limited means to prospectively assess these tools' fit for use. Through formative interviews with Wikipedia community members and AI developers, we derived a set of design requirements for systems that aim to support community-driven data curation. Based on these design requirements, we then developed Wikibench, a system that enables community members to collaboratively curate AI evaluation datasets, while navigating disagreements and ambiguities through discussion. As illustrated in Figure 1, community members can use Wikibench to select data points for inclusion in datasets, label data points with “individual labels” reflecting their personal judgments, and discuss their perspectives to decide upon a “primary label” for the data point. Through a field study on Wikipedia, we find that datasets curated using Wikibench can effectively capture community consensus, disagreement, and collective uncertainty. We demonstrate how Wikibench datasets can help in understanding areas of alignment and misalignment with community perspectives. Furthermore, we gain insight into the ways Wikipedia community members collaborate using Wikibench. We find that participants in our study used Wikibench to proactively shape the overall data curation process beyond labeling data, including refining label definitions, determining data inclusion criteria, and authoring data statements.

Overall, we demonstrate the potential of *community-driven* data curation, and contribute the following:

- **System**: We introduce Wikibench, the first system that supports community-driven curation of AI datasets.
- **Field study**: We present findings from a field study on Wikipedia to understand how Wikipedia community members interact with this system to collaboratively curate evaluation datasets.
- **Future directions**: Based on our findings, we propose future directions for HCI systems that support community-driven data curation within and beyond the context of Wikipedia.

Our full paper is publicly available (Kuo et al., 2024).

References

[Jo and Gebru2020] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

[Kuo et al.2024] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-driven data curation for ai evaluation on wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

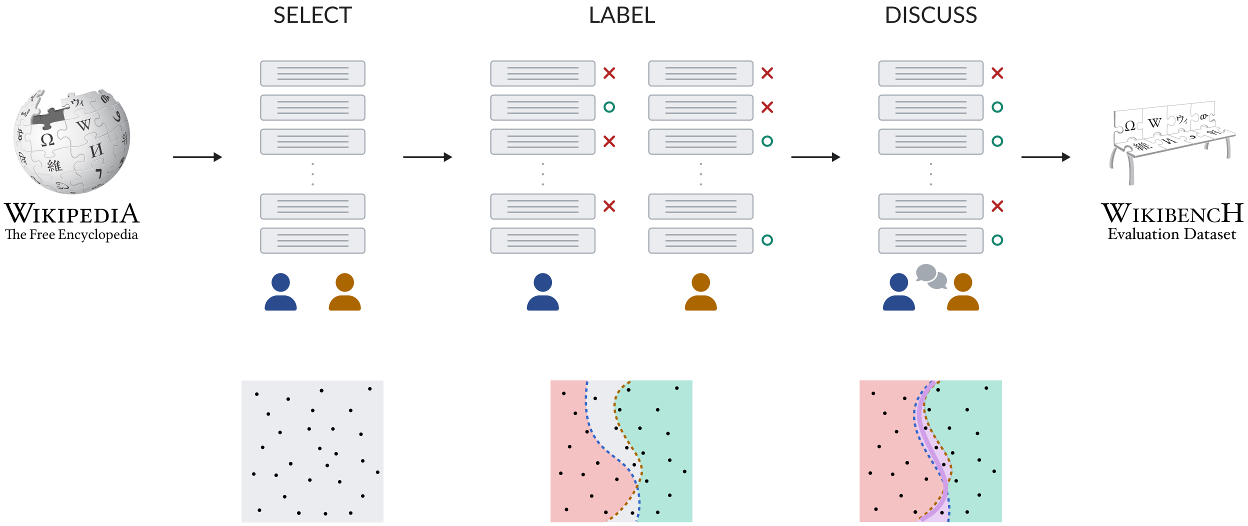

Figure 1: An overview of Wikibench's approach to supporting community-driven data curation. The top row illustrates community members' use of Wikibench to *select* data points (e.g., edits on Wikipedia) for inclusion in the dataset, *label* data points with "individual" labels based on their own initial judgments, and then *discuss* their perspectives and collectively decide on a "primary" label for the data point. The bottom row represents data points in a conceptual 2D space. As each community member labels data points, their labels form *decision boundaries* in aggregate (orange and blue dotted curves). Through discussion, participants may resolve some disagreements or clarify ambiguities in labeling, leading to changes in their individual labels. In addition, community members decide on a primary label for each data point, forming a consensus-based decision boundary (purple curve). Wikibench datasets preserve information about disagreement among community members (purple shaded region). The Wikipedia logo is licensed by Wikimedia Foundation, CC BY-SA 3.0, via Wikimedia Commons.