Community Literacy of Machine Learning: Engineering practical support for participatory design and auditing

Aaron Halfaker
Wikimedia Research
San Francisco, CA, USA
ahalfaker@wikimedia.org

Stuart Geiger
Univ. of California, Berkely
Berkeley Inst. of Data Science
Berkeley, CA, USA
sgeiger@gmail.com

 CSCW'18, November 3-7, 2018, Jersey City, NJ, USA.
https://doi.org/XXXXX/XXXXX

Abstract

Algorithmic systems—from rule-based bots to machine learning classifiers—have a long history of supporting the essential work of content moderation and other curation work in peer production projects. From counter-vandalism to task routing, basic machine prediction has allowed open knowledge projects like Wikipedia to scale to the largest encyclopedia in the world, while maintaining quality and consistency. However, conversations about what quality control should be and what role algorithms should play have generally been led by the expert engineers who have the skills and resources to develop and modify these complex algorithmic systems. In this paper, we breifly describe ORES: an algorithmic scoring service that supports real-time scoring of wiki edits using multiple independent classifiers trained on different datasets. ORES decouples three activities that have typically all been performed by engineers: choosing or curating training data, building models to serve predictions, and developing interfaces or automated agents that act on those predictions. This meta-algorithmic system was designed to open up socio-technical conversations about algorithmic systems in Wikipedia to a broader set of participants. We detail 2 key case studies that highlight the how Wikipedians come to understand and direct the use of ORES.

Author Keywords
Algorithm; Fairness; Transparency; Wikipedia; Collaboration; Participatory design; Auditing

ACM Classification Keywords
G.4 [MATHEMATICAL SOFTWARE]: Algorithm design and analysis; H.4.2 [Types of Systems]: Decision support (e.g., MIS); H.5.3 [Group and Organization Interfaces]: Collaborative computing

Introduction
Wikipedia—the free encyclopedia that anyone can edit—faces many challenges in maintaining the quality of its articles and sustaining the volunteer community of editors. The people behind the hundreds of different language versions of Wikipedia have long relied on automation, bots, expert systems, recommender systems, human-in-the-loop assisted tools, and machine learning to help moderate and manage content at massive scales. The issues around artificial intelligence in Wikipedia are as complex as those facing other large-scale user-generated content platforms like Facebook, Twitter, or YouTube, as well as traditional corporate and governmental organizations that must make and manage decisions at scale. And like in those organizations, Wikipedia's automated classifiers are raising new and old issues about truth, power, responsibility, openness, and representation.

Yet Wikipedia's approach to AI has long been different than in corporate or governmental contexts typically discussed in emerging fields like Fairness, Accountability, and Transparency in Machine Learning (FATML) or Critical Algorithms Studies (CAS). The volunteer community of editors has strong ideological principles of openness, decentralization, and consensus-based decision-making. The paid staff at the non-profit Wikimedia Foundation—which legally owns and operates the servers—are not tasked with making editorial decisions about content[1]. This is instead the responsibility of the volunteer community, where a self-selected set of developers build tools, bots, and advanced technologies in broad consultation with the community. Even though Wikipedia's longstanding socio-technical system of algorithmic governance is far more open, transparent, and accountable than most platforms operating at Wikipedia's scale, ORES[2], the system we present in this paper, pushes even further on the crucial issue of who is able to participate in the development and use of advanced technologies.

The politics of algorithms
Algorithmic systems play increasingly crucial roles in the governance of social processes[5]. Software algorithms are increasingly used in answering questions that have no single right answer and where prior human decisions used as training data can be problematic [1]. Algorithms designed to support work change people's work practices, shifting how, where, and by whom work is accomplished[2, 9]. Software algorithms gain political relevance on par with other process-mediating artifacts (e.g. laws[7]).

There are repeated calls to address power dynamics and bias through transparency and accountability of the algorithms that govern public life and access to resources[3, 8]. The field around effective transparency and accountability mechanisms is growing. We cannot fully address the scale of concerns in this rapidly shifting literature, but we find inspiration in Kroll et al's discussion of the potential and limitations of auditing and transparency [6] and Geiger's call to go "beyond opening up the black box[4]".

[1] Except in rare cases, such as content that violates U.S. law, see http://enwp.org/WP:OFFICE

[2] https://ores.wikimedia.org and http://enwp.org/:mw:ORES

We discuss a specific socio-political context—Wikipedia's algorithmic quality control and socialization practices—and the development of novel algorithmic systems for support of these processes. ORES implements a meta-algorithmic intervention aligned with Wikipedians' principles and practices: deploying a set of prediction algorithms as a service and leaving decisions about appropriation to the volunteer community. Instead of training the single best classifier and implementing it in our own designs, we embrace public auditing, re-interpretations, and appropriations of our models' predictions as an *intended* and *desired* outcome. Extensive work on technical and social ways to achieve fairness and accountability generally do not discuss this kind of socio-infrastructural intervention on communities of practice.

The ORES system

ORES has been iteratively engineered to meet the needs of Wikipedia editors and the tools that support their work. At the core, ORES is a collection of machine classifier models and an API. These models are designed and engineered by a varied set of model builders (some external researchers and others by our own engineering team) using varied sources of *training data*. The models that ORES hosts are engineered to support Wikipedian processes related to damage-detection, quality-assessment, and topic-routing, but the system is adaptable to a wide range of other models.

To make these models available for users, ORES implements a simple container service where the "container," referred to as a *ScoringModel*, represents a fully trained and tested prediction model. All *ScoringModels* contain metadata about when the model was train/tested and code for feature extraction. All predictions take the form of a JSON document. The ORES service provides access to ScoringModels via a RESTful HTTP interface and serves the predictions (JSON documents) to users. We chose this service structure because Wikimedian tool developers (our target audience) are familiar with this RESTful API/JSON workflow due to the dominant use of the MediaWiki API among tool developers. See section **??** for detailed examples of ORES' outputs and descriptions of how we engineered the system to support Wikipedians' work practices and the tools they use.

Case Study: Patrolling/ORES (Italian Wikipedia)

In this section, we describe a single case of collaborative auditing of ORES. While there are many cases like this one worth highlighting, the case of Italian Wikipedia is an excellent example of collaborative grounded theory deployed *in the field* to audit a machine prediction model.

Italian Wikipedia was one of the first wikis where we deployed basic edit quality models. Our local collaborator, who helped us develop the language specific features, User:Rotpunkt, created a page for ORES[3] with a section for reporting false-positives ("falsi positivi"). Within several hours, Rotpunkt and a few other editors noticed some trends. These editors began to collect false positives under different headers representing themes they were seeing. Through this process, editors from Italian Wikipedia were effectively performing an inductive, grounded theory-esque exploration ORES errors, trying to identify themes and patterns in the errors that ORES was making.

One of the themes they identified fell under the header: "corrections to the verb for *have*" ("correzioni verbo avere"). It turns out that the word "ha" in Italian translates to the English verb "to have". While in Englih and many other langauges, "ha" is laughing and adding "ha" repeatedly is a common type of vandalism seen in all langauges of

[3] https://it.wikipedia.org/wiki/Progetto:Patrolling/ORES

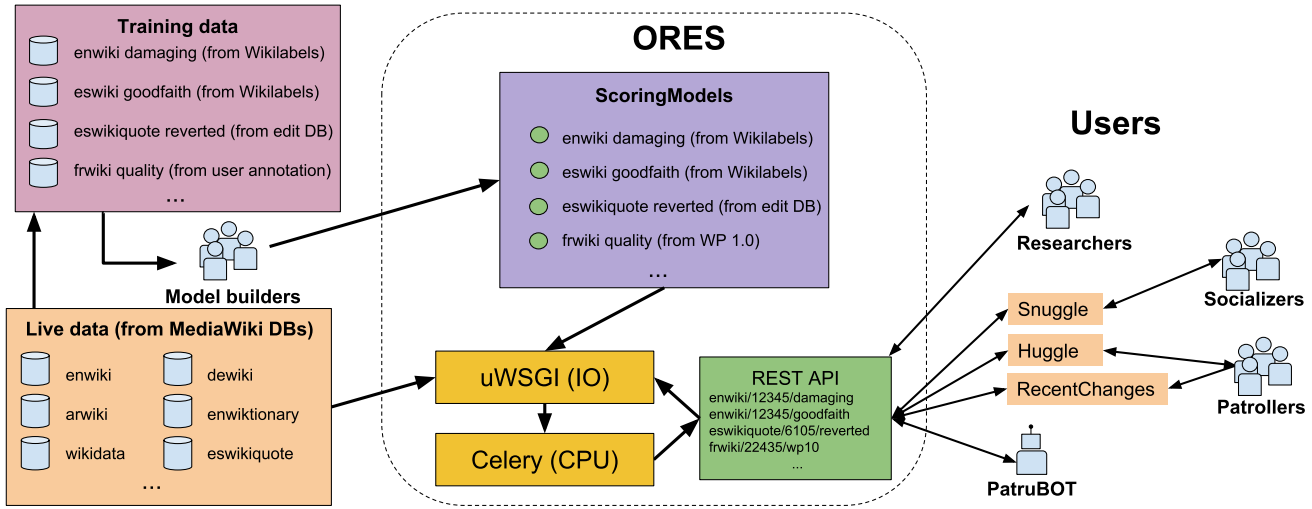

Figure 1: ORES conceptual overview. Model builders design process for training ScoringModels from training data. ORES hosts ScoringModels and makes them available to researchers and tool developers.

Wikipedia. We'd built a common feature in the damage model called "informal words" that captured these types of patterns. But in this case, it was clear that in Italian "ha" should not carry signal while "hahaha" still should.

Because of the work of Rotpunkt and his collaborators in Italian Wikipedia, we were able to recognize the source of this issue (a set of features intended to detect the use of *informal language* in articles) and to remove "ha" from that list for Italian Wikipedia.

Discussion

Deploying ORES as a meta-algorithmic probe allows us to ask new questions about how people understand the algorithms that govern (or at least supplement the governance of) their spaces. In the case above, it's clear that while many participants in this analysis could not approach technical jargon of model fitness (e.g. *precision* and *recall*), they were able to effectively evaluate the behavior of their machine prediction model and to detect problematic trends (biases) that the model expressed.

The case of Italian Wikipedia and the way that people operate around ORES should prompt us to think about machine learning literacy and power differently. This case shows that in one key respect (bias detection via crowd auditing), formal knowledge of machine learning was not necessary. Instead, the openness of ORES (the ability to deterministically get the same prediction for the same edit and share

that with others) made it possible for a large group of people to work together to build understanding.

REFERENCES

1. Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece. *SSRN. Paper presented at Governing Algorithms conference.* (2013). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322
2. Kate Crawford. 2016. Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41, 1 (2016), 77–92.
3. Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* 5, 7 (2017), 809–828. DOI: http://dx.doi.org/10.1080/21670811.2016.1208053
4. R. Stuart Geiger. 2017. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society* 4, 2 (2017), 2053951717730735. DOI: http://dx.doi.org/10.1177/2053951717730735
5. Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014).
6. Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.
7. Lawrence Lessig. 1999. *Code: And other laws of cyberspace*. Basic Books.
8. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.
9. Shoshana Zuboff. 1988. *In the age of the smart machine: The future of work and power*. Vol. 186. Basic books New York.