Enriching judgment as a new dimension in Wikipedia

Adam Wight
Wikimedia Foundation
San Francisco, CA, USA
awight@wikimedia.org

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

Abstract

We introduce a system for rich feedback in Wikipedia, called Judgment and Dialogue Engine (JADE). Expanding on current auditing approaches, JADE adds a dimension of human communication and collaborative decision-making between the auditors.

This rich feedback is currently targeted at machine learning models running on Wikipedia, which will benefit from the exploration and mitigation of biases. JADE makes our algorithms accountable to wiki users. Built on top of the collaborative practice and culture of Wikipedia, we anticipate that group judgments will offer a high-quality source of auditing data which can be used to counteract vicious cycles reinforcing the status quo.

In the wider context, our hope is that JADE will demonstrate a new technique to improve AI fairness and performance in general, and may help establish transparent and collaborative auditing as an appropriate intervention in many domains.

Author Keywords

Algorithm, Transparency, Collaboration, Wikipedia, Auditing, Machine learning

Introduction

We're building a new system for Wikipedia, to expand on an already impressive capacity for collective curation. This "judgment and dialogue engine" (JADE[1]) will give curatorial communities the tools and structure to allow for rich discussions, producing human- and machine-readable collaborative opinions that are appropriate for grounded theory research. Judgment data is expected to challenge our artificial intelligences in a way that reduces bias.

In today's context, the power dynamics of large-scale data collection and analysis are completely imbalanced, with key decisions made by unaccountable, corporate entities. The human subjects of data have become little more than rats in an experiment. Beyond the passively collected, digital traces of our lives, we rarely find an opportunity to participate explicitly in the cycle of data collection, analysis, and algorithm design. When our feedback is solicited, it's usually in a form completely stripped of agency, as basic as punching a happy or sad button.[3]

Wikipedia is an exceptional context, however, and one in which we already know that empowering users leads to a virtuous cycle of increasing quality and capacity. Giving the users powerful tools to collaboratively critique articles will likely bring about better articles, and will strengthen the reader-to-leader pipeline [4] in which users grow roots in their community, moving beyond individual efforts, to a stage where they form tightly connected groups who work together.

The main cultural productions on a wiki are "article" content about a subject and "talk" pages where the content or process is discussed. Routine anti-vandalism work involves making judgments about edits to the wiki content, and either marking the changes as "patrolled" (safe), or "reverting" the bad edits, removing them from content. Our initial goal is to enrich this type of activity, with new article and talk pages to encourage collaborative reflection about the edits. For example, an edit that adds an irrelevant link to an article could be marked in JADE as "damaging" and "spam", and the reviewer might note their suspicion that a large PR firm has been bankrolling similar vandalism in this topic area. Another reviewer could come along later and disagree, noting that the link is in fact helpful and giving their own justification.

We expect this type of exchange to be generative, and for the communities to exhibit emergent properties far beyond anything we've imagined. As the community evolves, we will try to adapt the software to better suit their needs.

Rich Feedback

One of the initial motivations for JADE was for users to make false positive reports against our "Objective Revision Evaluation Service" (ORES), a container for machine learning models running on Wikipedia data. A false positive in this case would be a prediction from ORES saying that an edit is vandalism, when in fact it isn't. These reports could be submitted with something as basic as a "right/wrong" button. However, there are multiple benefits to gathering rich feedback, both for the users and for the quality of data collected. The feedback process itself can lead users to gain a better understanding of and trust in our machine learning models. Simply asking for a freeform text note along with feedback leads to higher data quality. At the far extreme of rich feedback, the users can actually modify the model in real-time and examine the impact of their changes.[1] [8] On the machine side, we can provide rich explanations of why the algorithm made a given prediction, even breaking out the factors involved and allowing the user

[1] https://www.mediawiki.org/wiki/JADE

to annotate the factors directly.

We've decided to collect three types of rich feedback: freeform text, discussion, and structured values such as "damaging" in the above example. Wikipedia's cultural practices and software are already well-adapted to these. Other elements of rich feedback such as real-time, interactive manipulation of models will have to wait for a later iteration of JADE.

Collaborative Auditing

Another motivation for JADE is to help mitigate biases in ORES machine learning models. To illustrate our nightmare scenario, it's possible that Wikipedia editors would blindly follow ORES predictions and revert any new material the machine flags as potential vandalism. Such actions would reinforce whatever biases were present in the original training data, make the models less tolerant of content at the margins. The already Anglo-, Eurocentric norms[2] would become even more so.

The promise of JADE is that editors can discuss borderline and outlier cases, and help system designers identify the shortcomings in ORES. Editors are much more knowledgeable than us about their own local- or language-based issues, and have been generous enough to correct problems with our machine learning models. We can learn from the consistent patterns in their observations, or in the future, it might become possible for the editors to directly revise our machine learning models without any mediation by technicians.

By making a public audit of our algorithms' output, external researchers are able to make their own analyses, something that should be a right of "data subjects" worldwide. This is much like having an external financial audit. Even if nothing dysfunctional is happening behind the scenes, demonstrating a clean bill of health will improve understanding and trust in the system.[5]

Collaborative Judgments

A crucial detail of our system is that the judgments will be collaborative. This is more than just an aggregation of individual opinions. Group judgments arrived at through discussion are likely to be more accurate, can better estimate extreme values, and have a more realistic self-confidence level than any other method of aggregation.[7]

Heterogenous groups which disagree on their opinions seem to produce more accurate results, suggesting that a massive, public collaboration mechanism will be more successful than any small group of selected experts.[6]

Challenging Power

The reasons to audit powerful algorithms range from the unfair effects on individuals, to the meso-level health of each algorithm and its ability to deliver on its own organization's narrowly defined goals, to the health and survival of our society as a whole, at least to the degree that our social cohesion is now determined by technical algorithms. Many routes to auditing are tightly closed off by legal and economic forces. For example, Sandvig (2014) points out that the U.S. Computer Fraud and Abuse Act makes some of the most effective research methods illegal, and people have actually been prosecuted and imprisoned under this law. Companies have an economic incentive protect the internals of their algorithms as some of their most valued intellectual property, for example Google's much-speculated-about PageRank.

It's inconceivable that these companies will voluntarily offer any transparency. As Frederick Douglass said in 1857, "Power concedes nothing without a demand."

JADE hints at one potential weakness in the structure of the new algorithmic power, if public and transparent methods turn out to offer higher data quality and greater trust in the resulting products. In other words, the open culture community might be able to beat the commercial world at its own game.

There are some risks to this approach, of course. The open data sources that are developed may be exploited, with closed algorithms benefitting from our advances as we gain little from theirs. The mechanisms that we build may be adopted in commercial software, perhaps with modified and gamified incentives. Still, collaborative work seems to be the element of our system which is most resistant to cooptation: if individuals are alienated and their data extracted in isolation, they are by definition not building a collaborative judgment and we think the data quality will be lower. If they are truly collaborating, then they are by definition sharing knowledge and creating openness and transparency, at some scale.

Finally, as with any community-building project, we expect that a new power base will coalesce around JADE and that people involved will have their own ideas about what to do next. We hope that this community will shape future research, that we can support their evolving goals, and expand the scope of what they can accomplish. These groups will be at the cutting edge of a new socio-technological intervention and will have the shared practice of actually doing this work together.

REFERENCES

1. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
2. Heather Ford, Shilad Sen, David R Musicant, and Nathaniel Miller. 2013. Getting to the source: Where does Wikipedia get its information from?. In *Proceedings of the 9th international symposium on open collaboration*. ACM, 9.
3. Ville Levaniemi and Heikki Väänänen. 2012. Indicator of Satisfaction. (Aug. 16 2012). US Patent App. 13/502,668.
4. Jennifer Preece and Ben Shneiderman. 2009. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS transactions on human-computer interaction* 1, 1 (2009), 13–32.
5. 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.
6. Stefan Schulz-Hardt, Felix C Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. 2006. Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology* 91, 6 (2006), 1080.
7. Janet A Sniezek and Rebecca A Henry. 1989. Accuracy and confidence in group judgment. *Organizational behavior and human decision processes* 43, 1 (1989), 1–28.
8. Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.