Debiasing Vandalism Detection Models at Wikidata (Extended Abstract)

Presentation of work originally published in WWW '19 [He19]

Stefan Heindorf[1] Yan Scholten[2] Gregor Engels[3] Martin Potthast[4]

The ethics of artificial intelligence have become a major societal issue, evidenced also by them becoming a focus of attention of policy making. Recently, the European Union released "Ethics Guidelines for Trustworthy AI" — as did the IEEE and major companies such as Google, Microsoft, and IBM.[5] A central point in all the guidelines is the fairness of machine learning models and the mitigation of discrimination against minorities based on biased models. In this extended abstract, we report on a case study on debiasing vandalism detection models at Wikidata, the crowdsourced knowledge base of the Wikimedia Foundation.

Knowledge bases play an important role in modern information systems. For instance, web search engines use them to enrich search results, conversational agents to answer factual questions, and fake news detectors for fact-checking. Collecting knowledge at scale still heavily relies on crowdsourcing: Google acquired the open Freebase project to bootstrap its proprietary "Knowledge Graph" until Freebase was shut down and succeeded by Wikidata, the free knowledge base of Wikimedia. Other prominent open knowledge bases like YAGO and DBpedia also depend on crowdsourcing by extracting knowledge from Wikipedia. As crowdsourcing knowledge has a long history, so does the fight against damage caused by vandals and other users, which may propagate to information systems using the knowledge base, potentially reaching a wide audience.

Wikidata makes for an interesting case study to analyze and mitigate biases as it has one of the largest online communities and provides opportunities to pay particular attention to the content rather than the user reputation. Unfortunately, it is still common practice to identify malicious edits via meta data such as geolocation of IP addresses, age of user account, or language of edited content. While those features are simple to obtain, they do not directly judge the quality of an edit and harm well-intentioned users.

[1] Paderborn University, heindorf@uni-paderborn.de
[2] Paderborn University, yascho@mail.uni-paderborn.de
[3] Paderborn University, engels@uni-paderborn.de
[4] Leipzig University, martin.potthast@uni-leipzig.de
[5] https://ec.europa.eu/futurium/en/ai-alliance-consultation/; https://ethicsinaction.ieee.org/; https://ai.google/principles/; https://www.microsoft.com/en-us/ai/our-approach-to-ai; https://www.ibm.com/watson/ai-ethics/

In our research, we revealed for the first time that state-of-the-art vandalism detectors employed at Wikidata [He16, SHT17] are heavily biased against certain groups of editors. For example, *benign* edits of anonymous users receive vandalism scores over 300 times higher than *benign* edits of registered users. Such a widespread discrimination of certain user groups (especially that of anonymous editors) undermines the founding principles on which Wikimedia's projects are built.[6] Although the discrimination of anonymous users has long been recognized and the problem has been tackled through community outreach,[7] when discrimination gets encoded into automatic decision-making, this aggravates the problem. For example, it has been previously found that desirable newcomers whose edits are automatically reverted are much more likely to withdraw from the project [Ha13].

We carefully analyzed different sources of bias in Wikidata's damage control system and developed two new machine learning models that significantly reduce bias compared to the state-of-the-art. Our model FAIR-E uses graph embeddings to check the content's correctness without relying on biased user features. Our model FAIR-S selects the best-performing hand-engineered features under the constraint of no user features. Furthermore, we experiment with different transformations of the state-of-the-art vandalism detector WDVD: post-processing scores, reweighting training samples, and combining approaches via ensembles. We evaluate our approaches on a subset of the standardized, large-scale Wikidata Vandalism Detection Corpus 2016 [He17], and compare our results to others from the literature. Our best model FAIR-S reduces the bias ratio of WDVD from 310.7 to only 11.9, while maintaining high predictive performance at 0.963 ROC_{AUC} and 0.316 PR_{AUC}.

Keywords: Bias; Fairness; Knowledge Base; Wikidata; Data Quality; Vandalism

Acknowledgements: This work was supported by the German Research Foundation within the Collaborative Research Center "On-The-Fly Computing" (CRC 901).

Bibliography

[Ha13] Halfaker, A.; Geiger, R. S.; Morgan, J. T.; Riedl, J.: The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity is Causing Its Decline. American Behavioral Scientist, 57(5), 2013.

[He16] Heindorf, S.; Potthast, M.; Stein, B.; Engels, G.: Vandalism Detection in Wikidata. In: CIKM. 2016.

[He17] Heindorf, S.; Potthast, M.; Engels, G.; Stein, B.: Overview of the Wikidata Vandalism Detection Task at the WSDM Cup 2017. In: WSDM Cup. 2017.

[He19] Heindorf, S.; Scholten, Y.; Engels, G.; Potthast, M.: Debiasing Vandalism Detection Models at Wikidata. In: WWW. 2019.

[SHT17] Sarabadani, A.; Halfaker, A.; Taraborelli, D.: Building Automated Vandalism Detection Tools for Wikidata. In: WWW (Companion Volume). 2017.

[6] https://meta.wikimedia.org/wiki/Founding_principles

[7] https://en.wikipedia.org/wiki/Wikipedia:IPs_are_human_too