Automated detection of Wiki Misconduct!
credit for this project to
Arnab Sarka  Charu Rawat  Sameer Singh
2019 graduates, Master of Data Science
Data Science Institute, University of Virginia
project page on Meta-Wiki

presentation by
Lane Rasberry / bluerasberry
Wikimedian in Residence
Data Science @ U of Virginia

more credits at end!
Goal of this project

1. Use machine learning to examine all blocked users in English Wikipedia.
2. Based on the activity of blocked users, produce an algorithm which can rank users by the amount of likely misconduct they do.
3. Have the police bot patrol Wikipedia and publish a list users which it accuses of misconduct.
1. Machine learning for robot moderation
2. Automated wiki moderation
3. How to go further
4. How to do research in wiki
Machine learning for moderation

1. humans do moderation millions of times in an online platform
2. artificial intelligence (AI) examines that moderation and creates an algorithm to replicate it
3. the artificial intelligence does labor using that algorithm
4. humans oversee the AI to correct it
5. soon, the AI needs less human support to do work
Want to learn more?

- machine learning
- data science
- artificial intelligence
- text mining
- natural language processing
Automated wiki moderation

Woooooo view the publications!
Media from the project

- research article
- on-wiki reporting
- ethics review
- video report
Automated wiki moderation

get data
work for 8 months
publish in the open
next cycle
Summary

1. collect list of English Wikipedia's 1.1 million blocked users, 2004-2018
2. consider 6 million unblocked users, 2017-18
3. too much data to compute, exclude 95%
4. apply machine learning models
5. generate list of users who likely merit a block
Going further

Pros
1. greatly aids humans
2. scales
3. inexpensive
4. recruits new community
5. really cool

Cons
1. replaces humans
2. propagates wildly
3. organization required
4. new disparities
5. creepy af
Recommendations for next steps

1. promote wiki discourse on human / bot interaction
2. standard labels in wiki moderation
3. develop wiki research infrastructure
4. establish public / private spaces to review misconduct accusations ethically
How to do wiki research

1. document on wiki
2. free and open licensing
3. leave the wiki community members ALONE
4. publish self-evaluation of ethics
5. share credit with wiki community organizations
Thanks and credits

U of Virginia research team
U of Virginia faculty advisor
Wikimedia Foundation Trust and Safety
Wikimedia New York City
Wikimedia LGBT+
Wikimedia Medicine
Wikimedia Tool Labs
Wikimedia Foundation Grants
SlidesCarnival

Charu, Arnav, Sameer
Raf Alvarado
S. Poore, P. Early, C. Lo
for Code of conduct project
for misconduct consulting
for spam consulting
for online database access
for US$5000 grant
slides template
user:bluerasberry