Using Digital Tools to Tackle Online Harms and Extremism

23 Nov 2023

Zeerak Talat <u>z@zeerak.org</u> | @zeeraktalat www: <u>zeerak.org</u>

Agenda

- Quick deep dive into AI for language technology
- Identify and discuss some weaknesses of AI
- Practical methods for inclusion in your work
- Defining and finding useful errors

r language technology weaknesses of Al Jsion in your work Jl errors

Natural Language Processing (NLP)

03

Operates on a foundational assumption: The Distributional Hypothesis

- Semantic (i.e., literal) meaning can arise from word frequencies

AI for Classification

- 1. Develop language model
- 2. Create training data
- 3. Further develop language model with training data 4. Use the resulting model to classify new data

AI for Classification

- 1. Develop language model
- 2. Create training data
- 3. Further develop language model with training data
- 4. Use the resulting model to classify new data

TIME

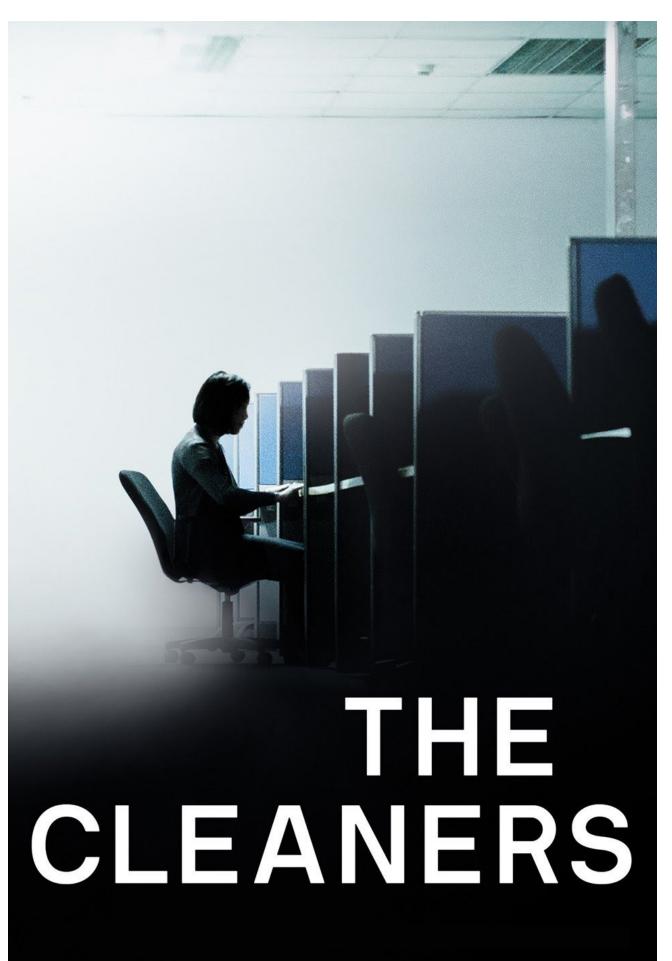
Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

AI for Content Moderation

1. Existing models developed for general purpose 2. Training data difficult to get right

AI for Content Moderation

1. Existing models developed for general 2. Training data difficult to get right



06

Online Harms and Extremism



A person whose reflection is being distorted by mirrors. Source: funplanners.com

Working with AI

07

- 1. Developing your data for your needs
- 2. Identifying crucial failure modes for your case
- 3. Attend to the kind of errors that are most consequential

Recommendations and Conclusion

- Develop own data with experts
- Identify the particularly weaknesses of data
- Assume AI will be wrong in useful ways
 - When identifying a particular error, look for more of them

08