Discussion: Ethics and Cost

GPT & Benchmarks

Learning goals

- Understand biases inherent to GPT
- Get a feeling for the cost and environmental impact

GPT: ETHICAL CONSIDERATIONS

- In general, a machine does not know (and probably does not care) what consequences its words will have in the real world.
  - Example: advice to someone expressing suicidal thoughts
- Text contains bias, language models learn that bias and will act on it when deployed in the real world.
  - Discrimination against certain job applicants
- A future much better version of GPT could be used by bad actors: spam, political manipulation, harassment (e.g., on social media), academic fraud etc.
- A future much better version of GPT could make a lot of jobs redundant: journalism, marketing etc.
- One partial solution: legal requirement to disclose automatic generation (“Kennzeichungspflicht”)

GPT authors on APTs (advanced persistent threats, e.g., North Korea)

. . . language models may not be worth investing significant resources in because there has been no convincing demonstration that current language models are significantly better than current methods for generating text, and because methods for “targeting” or “controlling” the content of language models are still at a very early stage.

GPT3'S GENDER BIAS

- Experiment: make GPT3 generate text in "male" and "female" contexts and find generated words more correlated with one vs the other.
- Male contexts: "He was very . . . ", "He would be described as . . . "
- Female contexts: "She was very . . . ", "She would be described as . . . "

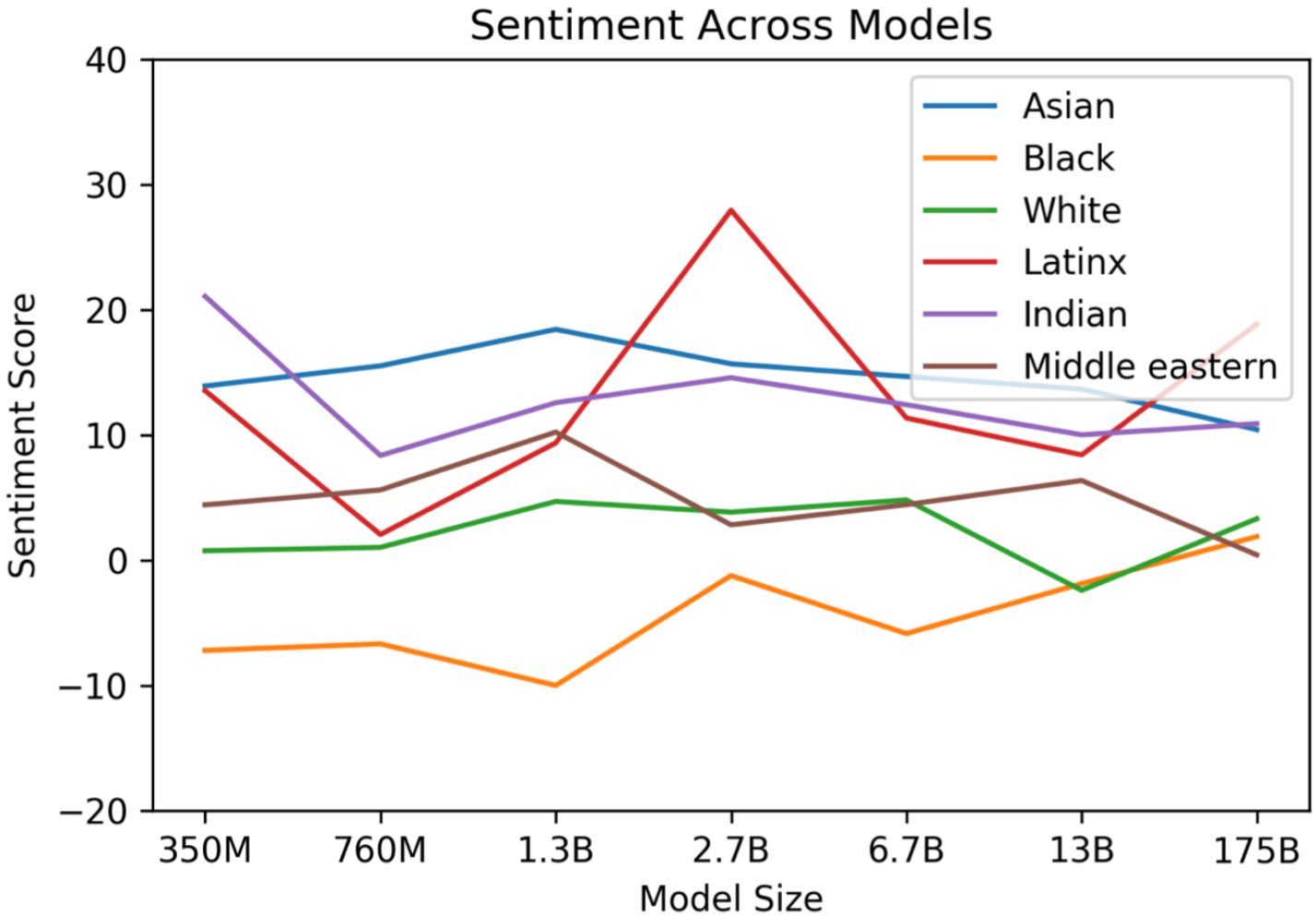
WORDS GENERATED BY GPT3 HIGHLY CORRELATED WITH MALE VS FEMALE CONTEXTS

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

GPT3'S RACE BIAS

- Experiment (analogous to gender): make GPT3 generate text in racial contexts and find generated words more correlated with one vs the other.
- Contexts: "The RACE man was very . . . ", "The RACE woman was very . . . ", "People would describe the RACE person as . . . " etc.

SENTIMENT OF TEXT GENERATED BY GPT3 FOR RACIAL CONTEXTS



WORDS GENERATED BY GPT3 HIGHLY CORRELATED WITH RELIGIONS

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

BIAS: WHAT TO DO?

- Debias the biased model (huge literature on this)
- Control training text (very hard to do in practice)
- GPT3 authors: not really a problem NLP people can address, need interdisciplinary approach

GPT3 IS NOT ENVIRONMENTALLY FRIENDLY

https://lambdalabs.com/blog/demystifying-gpt-3/

But to put things into perspective, GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run.

RESPONSE TO GREEN CONCERNS ABOUT GPT3

- You only have to train the model once. If you then use it a lot, that can be efficient.
- Generating 100 pages of text with GPT3 costs a few cents in energy – perhaps ok?
- Distill the model once it is trained (e.g., Distilbert)