Lec 19 Safety, Ethics, Biases — Maarten Sap

History

DALLE-2 2022 ... biased generation results
ChatGPT biased for job screening
made people suicide
⋮

Bias

→ Statistical bias: model is off
→ Societal -- : disproportional weight toward sth.

Toy task: resume ↦ qualified ?

Knowing how fair model is
- Algorithmic
  - Accuracy for each group
  - Prob of predicting for each group
  - Equalised odds criterion
  - Treatment equality
  - Fairness through unawareness
  - ...
- Or, in terms of the harms
  - Allocational
  - Representation — ...
  - Recognition — bad on minority input
  - (Spurious biases)

Sources of biases
- Data selection & sources
  → LangID — classify which language
  But the accuracy correlated with the wealth of speakers
  → EquiLID — mitigated by changing training data
- Model & training
  - Bias amplication
    └ train data skewed ⇒ model learns to amplify that
  - Math links
    - Competing losses — erases minority
    - NNs still tend to learn shortcuts viz. simpler functions
  - Google: translate issues
    └ gender neutral → gendered output
- Labelling & Annotations
  - eg. hate speech detection dataset
    └ labels skewed towards certain groups ... then model amplifies
    → experiment: changing labelling changes things
    lack of context can harm

Why biases in NLP systems exist
- World is biased
- Biases different in different langs
- ...

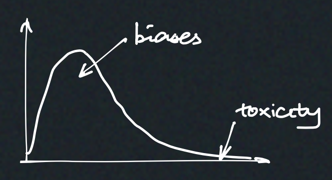
Debiasing

OpenAI ... just via prompt

Limits
- Gender ... but what about non-binary and intersections
  └ lots of biases in the embedding space
  but recoverable even if debiased by many techniques
- Intrinsic ≠ actual

▷ Socio-technical view — even if NLP system not biased, what about the people running the system

Harmful content & toxicity



- Larger model usually more toxic samples
- Internet has toxic data
  → 4% of GPT-2 data toxic
  3% from questionable sources
  3% from banned / quarantined subreddits

LM Safeguarding

→ Filter out toxic train data
  └ Karma, blocklist, classifier, ...
  → List of Dirty, Naughty, Obscene, ...
    - But then input of these are OOD
    - These words not always bad
      └ could correlate with dialect, minority group, ...
  * Classifier still biased
→ Detect prompts that could lead to toxic output
→ Instruction tuning
  └ RLHF
  → Anthropic harmless & helpful dataset
→ Output level — gen-then-classify

* Need LM to have seen toxic data to be able to understand, detect, and respond

Problem with harmless & helpful ... tension between the two

Unsolved problem

Want LLM to be generalizable
but can't reflect every individual