Chapter 6 - Why Can't AI Fix Social Media

TLDR

Content moderation is critical for the success of social media companies. Social media companies use a mixture of small machine learning models, fingerprinting and human moderators. There are a large number of content moderation failures ranging from embarrassing to utterly horrendous.
Key difficulties:

- Huge scale (number of pieces of content) results in small simplistic models that don't understand nuance
- Huge diversity of content (language, culture, topics) leads to an explosion of edge cases that need to be approached and handled differently
- Adversarial domain in which content produces adapt an evolve making machine learning (any system honestly) difficult

Personal take

This chapter was my favourite so far! The authors seem to have a lot of insight. Is this the question for our age? How do we do solve content moderation before the world implodes? What systems can we imagine working here? It seems particularly topical following the US election of Donald Trump.

When Everything Is Taken Out of Context

- Community is key asset to a social media company (and community requires content moderation)
- Content moderation already uses simple automated approaches (spam classifiers are good, other categories are not so require human review)
- Some example failures:
    - Swollen child genitals resulted in a account ban
    - Nazi punch cartoon
    - Rohingya genocide / ethnic cleansing in Myanmar potentially the most egregious example (seems like there was very limited content moderation, potentially due to lack of automatic translation, language expertise, prioritisation within Facebook)
- Failures in traditional machine learning approaches due to:
    - Data volume

    - Data quality
    - Computational resources (scale or posts, ad revenue model)
- Modern LLMs (GPT) may be better
- Humans are able to intuitively discern the cases however human moderates are required to follow overly prescriptive rules

Cultural Incompetence

- Traditional AI (scalable) algorithms struggle to take into account culture nuance
- Different cultures require different approaches / standards to content moderation
    - Societies without free press may not be robust to free speech (slightly ironic that America was mentioned as a counter example to this)

AI Excels at Predicting... the Past

- Speech (language) and the content of the speech evolves over time (e.g. slurs, covid conspiracy theory)
- Can AI ever assess the ground truth of a new claim?

When AI Goes Up against Human Ingenuity

- Content moderation is an "adversarial" problem
    - AKA content producers learn and evolve in response to content moderation, for example obfuscating speech ("goal weights", "safe and effective")
    - This is a notoriously difficult machine learning regime

A Matter of Life and Death

- Suicide is a case where social media may have an edge over medical profession due to more real time data
- Even low false positive rates may be harmful due to false reports having consequences (e.g. stigma, laws in some countries, low quality medical interventions)