CSE 291/DSC 291
Information Manipulation: Trustworthiness of Information in Cyberspace

Molly Roberts & Stefan Savage, Winter 2023

Content-based Interventions
Key idea

- If we can identify misinformation content then we can treat it differently and thus prevent/mitigate its spread/internalization

- A few questions emerge
  - How would you identify it?
  - How might you treat it differently?
  - Does that work?
  - Is it practical?
Drivers to moderate: carrots and sticks

- Attractiveness of platform (carrots)
  - Vibrant network of participants (free speech)
  - Civil discourse (manage trolls, hate speech, obscenity, etc)
  - Attractiveness to advertisers

- Regulatory pressures (sticks)
  - Child sexual-abuse materials (everywhere)
  - Hate speech in EU, Holocaust denial and antisemitism in Germany, Nazi symbols in France,
  - Sympathy with North Korea in South Korea
  - Etc...
Quick legal aside in US

- In early 90s some of the first commercial online services emerge – leading to claims of defamation
  - Compuserve – basically no moderation
    Cubby v Compuserve (1991) – court extends theory of conduit liability, distributor theory to this situation
    (Compuserve not a publisher unless they had reason to know)
  - Prodigy – family friendly, lots of moderation
    Stratton Oakmont v Prodigy Service (1995) – court says “nope, you exerted editorial control, so you’re a publisher” (simplified)
  - What incentives does this create?

- Stratton decision leads Wyman and Cox to introduce the Freedom and Family Empowerment Act, which ultimately becomes Section 230
Quick legal aside in US

- Section 230 of Communications Decency Act
  - (1) Treatment of Publisher or Speaker
    No provider or user of an interactive computer service shall be treated as
    the publisher or speaker of any information provided by
    another information content provider.
  - (2) Civil Liability
    No provider or user of an interactive computer service shall be held liable
    on account of—
    (A) any action voluntarily taken in good faith to restrict access to or availability of material that
    the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent,
    harassing, or otherwise objectionable, whether or not such material is constitutionally
    protected; or
    (B) any action taken to enable or make available to information content providers or others
    the technical means to restrict access to material described in paragraph (1).[1]

- Carve outs for federal criminal law, copyright and sex trafficking
Quick legal aside in US

- Section 230 of Communications Decency Act

- Effective meaning
  - Platforms can’t be held civilly liable for allowing third-party speech on their platforms
  - Platforms can’t be held civilly liable for censoring third-party speech on their platforms
  - Note this is only true for *third-party speech*
Human moderation: pros/cons

When might this work online?
Algorithmic moderation

- Basic idea: automate moderation work by analyzing the content
  - Classify content for various kinds of “unwantedness” and trigger if over some threshold

- Side question: is this an adversarial game?
  - How much are actors trying to evade moderation?
Algorithmic moderation in an old (adversarial) setting: spam filtering

- Blocklist of fixed words (e.g., Viagra, Cialis, Rolex, replica, free etc)
  - Response: encode content – Via graft or ViAGrA or V1agra

- Bayesian classification $P_r(spam|word) = \frac{P_r(word|spam)P_r(spam)}{P_r(word)}$
  - Response: add lots of extraneous text to spam to “dilute”
  - Poison training data
  - Send image instead of text

- OCR images
  - Response: Obfuscate image

- Focus on URLs (where does e-mail send you?)
  - Response: ask user to type in URL
Kinds of content identification

- Paper describes two broad approaches:
  - Matching
  - Classification

- Are these different in any important way?
Matching approaches

- Assumption is that there is a corpus of “known-bad” material
  - Copyright: MPAA movies, RIAA for audio
  - CSAM: NCMEC corpus + Microsoft PhotoDNA match
  - GIFCT: SIHD database + PhotoDNA

- Match against corpus
  - Corpus is big so optimize via hashing
  - But precise hashing can be problematic...

- Perceptual hashing
  - Match “if similar enough”
Example: PhotoDNA

- Tries to be robust to
  - Format transformation (e.g., JPG, GIF, PNG, etc)
  - Cropping
  - Color/brightness shift
  - Rotation and scaling
What are key pros/cons of matching?
Classification approaches

- Examples of wanted vs unwanted behavior in content
  - Aside: how might you get these?

- Train classifier to separate between classes using features in data
  - Can be used for text, images, audio, etc (also use content meta-data)
  - Common approach for assessing “toxicity” in text

- What is the tradeoff here vs matching?
In practice everything gets combined

- **Matching**
  - For known-bad stuff
  - Feature extraction: e.g., corpus of slurs may be treated specially

- **Classification**
  - Rank content and confidence for variety of dimensions (collection of classifiers)

- **Human moderation**
  - Classifications without high confidence can be punted to people

- **Crowdsourced**
  - People flag content they think violates policies
Quick aside: public vs private

- Platform decision about when to deploy these technologies
  - On public data (e.g., posts on Facebook)
  - On private data (e.g., private photo on Dropbox)
  - On shared data (e.g., messages you are sending on WhatsApp)

- Very different policies and different for different kinds of tasks
Ok, you’ve identified unwanted content… what to do?

- **Nudge**
  - Suggest user not post that thing just the way they did it

- **Deletion**
  - Content
  - User
  - Before posted? Or after?

- **Flagging**
  - Deprioritize content
    - Less likely to be shown (aka shadow ban)
    - Don’t put ads there (demonetize, also keep advertisers happy)
  - Send for further human review
The technical problem with algorithmic content classification

- Sources of error
  - Systems don’t understand context; context can be quite complex
  - Limited training data
  - Fundamental ambiguity
  - Trolls work around rules – hard to make robust classifiers

- Accuracy: False positives and false negatives
  - Act on content that you shouldn’t
  - Don’t’ act on content that you should

- 99% accuracy sounds good, but not at scale
  - 350B photos/day messages/day = 3.5 BILLION misclassifications/day
Aside – just definitions are hard

- Hate speech – what is it?
- When might it be illegal?
- What do you think Facebook’s line is?
  - *We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.*

  - *We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they’re referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for PC groups.*
Aside – just definitions are hard

- Hate speech – what is it?
- When might it be illegal?
- What do you think Facebook’s line is?
  - We also prohibit the usage of slurs that are used to attack people on the basis of their protected characteristics. However, we recognize that people sometimes share content that includes slurs or someone else’s hate speech to condemn it or raise awareness. In other cases, speech, including slurs, that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.
Extra-technical issues

▪ Transparency
  – What is process?
  – Can we explain decisions?

▪ Justice/Bias/Fairness
  – “All white people are racist. Start from this reference point, or you’ve already failed,” – Didi Delgado, BLM activist
  – Female Drivers, Black children, white men -- FB training document, which category is protected?

▪ Politicization
  – Moderation is powerful tool; a proxy for political struggles here and abroad