Anti-social Computing

13 December 2024 22:02

## Anti-Social Behavior

- Cluster B of personality disorders → behavior that deviates from expectations of individual's culture.
- ASB are → emotional, erratic, dramatic.
  - do → violate social norms
    - lying + deceit
    - impulsive
    - irritable
    - aggressive
    - lack of remorse
    - consistently irresponsible
    - disregard for safety

  } an individual w/ ASB personality can commit an act of harm + not feel guilt.

  ↓

  Anti-social computing is the study of how to explain + predict ASB.

  ↓

  ASB manifests as:
  - → trolling
  - → bullying
  - → flaming

- But why ASB?
  - → environmental
  - → genes
  - → neural factors.

## Trolling

- write offensive + inflammatory comments → sexist, racist, hateful, profane
- ♂ more likely than ♀
- aim: DISRUPT online discussion + GRAB ATTENTION → focus on small no. of threads + issues of petty things
  - ↓ become less tolerant
  - ↓ get kicked
- disregard author, no respect to other commenters.
  - → subtle troll
  - → threats + abuse
  - → but trolls can't write well
  - → whatever they write is irrelevant
- enjoy at the cost of others.

↓

Exposure to trolling = anxiety, depression } psychopathological outcomes.

## Cyberbullying

- TARGET individuals + cause them distress.
- post abusive + vicious comments about a SINGLE kid.
  - → intentionally + repeatedly harass by:
    - intimidating
    - shaming
    - demeaning
- They don't want attention

## Flaming

- post insults + offensive language online — fuelled by anonymity
- intensify conflict in controversial discussions
- causes division + drives users away.
- direct aggressive insult aimed at individuals + ideas (trolling is about baiting rxns)

- direct aggressive insult aimed at individuals + ideas (trolling is about baiting rxns)

## Dark Content

- online material that's harmful, disturbing, or malicious.
- exploit vulnerability
- spread fear + confusion

} Effect:
- content moderation challenges
- undermine community experiences
- create unsafe environ
- contribute to ASB.

## ASB Categories

### 1 – failure to conform to social norms.
- taboo words + words about law + legal system
- 1st + 3rd person pronouns
- express grudges towards legal system
- (that's why stop words are impt).

### 2. – irritability + aggressiveness.
- highest no. of aggressive, abusive, angry words
- use abusive words to convey anger.
- all swear words.

### 3 – reckless disregard for safety.
- big deviation from prev. classes.
- no abusive or taboo words.
- words of harm (kill, stab, murder) + fun words (fun, like)
- people have fun at expense of their safety + safety of others.

### 4 – lack of remorse
- lack of regret after hurting someone
- terms are both -ve + +ve
- -ve words are representatives of having done something wrong.
  +ve words represent indifference + discord after hurting.

### 0 – non anti-social
- nice, non aggressive, non-abusive.
- easy to identify w/ high accuracy.

## RL Setup – frame env. as finite markov decision process:

↓

Agent learns thro' trial + error

- finite set of states S.
- finite set of actions A
- transition probabilities T
- scalar rewards R for each transition
- discount factor γ for importance b/w short + long term rewards.

- discount factor $\gamma$ for importance b/w short & long term rewards.

↓

**Twitter MDP:**

action: share

feedback: re-share by others.

**Actions by agents:**
- make original content: active tweet: tw
- reshare others' content: active retweet: rt
- interact via reply/mention: active reply: rp
- keep silent: active nothing: nt

**Feedback from env:**
- reshare agent's tweet: passive retweet: RT
- interact w/ agent: passive reply: RP
- don't engage w/ agent: passive nothing: NT

## Inverse RL

↓

Agent given reward $f^n$ to achieve goal —— find reward that explains observed behavior

for every account, IRL gets 12 scalar values related to state-action
high reward in (s,a) action pair
→ user is v. motivated to perform action a in state s.

this approach tries to understand motivations & incentives b/w users & bots
by using these 12 reward values. (-ve) (+ve)

## Content Moderation Systems

① Content classification:
- categorize content into predefined categories based off platform's policies
- also enforce lang. & formatting requirements.

↳ overall content acceptability.

② Content annotation:
- specific words or phrases in content
- identify & flag bad lang.
- annotation examines individual elements for compliance.

③ Challenges:
- sparsity
- adversarial behavior
- asymmetric costs
- bias.