Class Reading Assignment 6: Tackling Harms of Social Computing Systems

Course: CS 6474 / CS 4803 Social Computing

Grade: 4% of overall course grade (40 points total)

Due Date: the last class of instructional period on April 27 | *Earlier submissions encouraged | Earlier submissions will be graded sooner and feedback provided ~2-3 weeks from the time of submission*
What to hand in: Submit as a single PDF on Canvas

Formatting Guidelines:
• Length: approximately 3 pages single-spaced, 1-inch margins
• Font: at least 11pt, readable serif or sans-serif

Grading Emphasis:

- Use of specific evidence from both papers (not generic descriptions)
- Accuracy in describing key findings and outcomes
- Clear comparison and synthesis across the two studies
- Direct, concise responses that address each prompt
- Evidence of reasoning beyond summary (e.g., explaining why outcomes occur)

Collaboration Policy:
This is an individual assignment. You may discuss high-level ideas with classmates, but all submitted work must be your own. You may not share written responses.

AI Use Policy:
You may use AI-based tools only for proofreading or improving clarity. You **may not** use AI tools to generate ideas, arguments, or structure. Responses should reflect your own reasoning and engagement with the readings and lectures.

This assignment builds directly on Week 12's lecture and discussions on "Tackling Harms: Online Content Moderation". Papers included:

- You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech [pdf]
- #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities [pdf]

Question 1: Effects of Content Moderation Strategies

Drawing on the Instagram (Chancellor et al.) study and the Reddit banning (Chandrasekharan et al) study, answer the following:

(a) (6 pts) Briefly describe one key outcome of content moderation in each study:

- Instagram (pro-ED communities)
- Reddit (banned subreddits)

Be specific about *what changed* in user behavior or content.

(b) (6 pts) Compare the effectiveness of moderation in the two settings. In what ways did moderation *work* or *fail* in each case? Use one concrete example from each paper.

(c) (8 pts) Based on the findings of both papers, which moderation approach appears more effective in reducing harmful behavior on-platform**,** and why? Support your answer using one piece of evidence from each study (e.g., increased engagement on variant tags vs. decrease in hate speech after bans). Limit your response to 2–3 sentences.

Question 2: Adaptation and Unintended Consequences

This question examines how users and communities respond to content moderation.

(a) (6 pts) In the Instagram study, describe how users adapted to content moderation through lexical variation. What are lexical variants? Provide one concrete example of how a moderated tag changed. Briefly explain how this helped users circumvent moderation**.**

(b) (6 pts) In the Reddit study, describe how users responded to subreddit bans.

- What happened to user activity levels after the ban?
- What happened to hate speech usage among users who remained?

Answer using specific findings from the paper.

(c) (8 pts) Based on both studies, identify one unintended consequence of content moderation. Clearly state the consequence. Briefly explain why it occurs, using evidence from one or both papers. *(Examples may include evasion, increased toxicity, user exit, migration—but must be justified using the readings.)*