Mitigating the Harm of Recommender Systems

Saloua Daouki

2025-06-29

Introduction

Recommender systems have become central to how information and media are consumed online. While they help users discover content, research and journalism have highlighted serious consequences, such as radicalization, misinformation, and algorithmic discrimination. This discussion draws on the following readings:

- Renee Diresta, Wired.com (2018): *Up Next: A Better Recommendation System*
- Zeynep Tufekci, The New York Times (2018): *YouTube, the Great Radicalizer*
- Sanjay Krishnan et al.: *Social Influence Bias in Recommender Systems: A Methodology for Learning, Analyzing, and Mitigating Bias in Ratings*

Harms of Recommender Systems

1. Radicalization and Echo Chambers

- **Tufekci (NYT, 2018)** describes how YouTube's recommendation algorithm, optimized for engagement, tends to push users towards more extreme and sensational content. This can create "rabbit holes" where users are exposed to increasingly radical ideas, regardless of their starting point.
- The platform's design unintentionally amplifies divisive, conspiratorial, or radical content because such material often drives more watch time and clicks.

2. Algorithmic Discrimination & Social Influence Bias

- **Krishnan et al.** discuss how recommender systems can amplify biases present in user ratings or social signals. For example, if early ratings are biased, subsequent users may be influenced by them, reinforcing and amplifying the bias ("social influence bias").
- Discrimination can emerge if the system systematically under-represents or misrepresents content from marginalized groups.

3. Filter Bubbles

- By tailoring recommendations to prior behavior, systems can limit exposure to diverse viewpoints ("filter bubbles"). This can reinforce existing beliefs and reduce the diversity of content encountered.

Mitigation Strategies

A. Algorithmic Interventions

1. Diversification & Serendipity

- Incorporate diversity and serendipity metrics into the recommendation objective, not just accuracy or engagement.

- Intentionally include content outside of the user's usual preferences to widen their perspective and break echo chambers.
- Example: Mix in a certain percentage of recommendations that are popular across different user groups, or that provide counterpoints to previously consumed content.

2. Debiasing and Fairness-aware Algorithms

- Use algorithms that explicitly measure and correct for bias in input data or recommendations.
- Implement fairness constraints (e.g., ensuring proportional representation of different groups or topics).
- Krishnan et al.'s methodology includes detecting and adjusting for social influence bias in ratings.

B. Human Oversight & Transparency

- Allow users to understand why something was recommended (explainability).
- Provide users with controls to adjust their recommendation preferences or filter settings.
- Regularly audit algorithms for harmful outcomes, using independent or external reviewers.

C. Product & Policy Design

- Limit autoplay or endless scroll features that encourage binge-watching/consumption.
- Provide warnings or educational prompts when users are being directed toward potentially extreme content.
- Collaborate with subject-matter experts to identify dangerous or misleading content patterns.

Relating to My Project 3 and Project 4 Work

In both my Project 3 and Project 4 assignments, I explored not just accuracy but also the diversity of recommendations generated by different algorithms on the MovieLens dataset. In Project 3, when using the SVD (matrix factorization) model with 5-fold cross-validation, I observed that all five users received the exact same top-5 recommended movies. Similarly, in Project 4, when I calculated intra-list diversity—using a function to quantify how varied the recommended items were for each user—I found that the diversity metric values were consistently zero. This indicated that the recommendation lists lacked variety and tended to suggest very similar or even identical items to different users.

Below is the function I used in Project 4 to calculate intra-list diversity:

```
calc_diversity <- function(topNList, sim_matrix) {
  lists <- as(topNList, "list")
  diversities <- sapply(lists, function(items) {
    if(length(items) <= 1) return(NA)
    sim_values <- sim_matrix[items, items]
    sim_values[lower.tri(sim_values, diag = TRUE)] <- NA
    mean(1 - sim_values, na.rm = TRUE)
  })
  mean(diversities, na.rm = TRUE)
}
```

This outcome is not desirable for a real-world recommender system, as high diversity is important for exposing users to a broader range of content and for mitigating issues like filter bubbles and radicalization, as discussed in the readings. The low (or zero) diversity scores I observed likely resulted from data sparsity and the limitations of the tested algorithms under the chosen evaluation scheme. This experience highlights a real challenge: while measuring and optimizing for diversity is a promising strategy to counteract some algorithmic

harms, practical implementation may require denser data, alternative algorithms, or different evaluation approaches to achieve meaningful increases in diversity.

Furthermore, although I also implemented a serendipity metric in Project 4 to encourage surprising and relevant recommendations, the limitations of my system similarly affected this metric. This illustrates the complexity of operationalizing fairness and diversity goals in real-world recommender systems, as emphasized in this week's readings.

Discussion

Recommender systems are not value-neutral; their design choices have real-world social consequences. The readings emphasize that algorithms optimized for engagement alone can inadvertently promote extremism or bias. Mitigation requires:

- Rethinking optimization goals,
- Incorporating fairness and diversity,
- Human-in-the-loop oversight,
- And, when necessary, regulatory or policy guardrails.

No single intervention will solve all problems, but a multi-pronged approach can help reduce the harms discussed.

References

- Diresta, R. (2018). Up Next: A Better Recommendation System. *Wired*. Link
- Tufekci, Z. (2018). YouTube, the Great Radicalizer. *The New York Times*. Link
- Krishnan, S., Patel, J., Franklin, M. J., & Goldberg, K. (n/a). Social Influence Bias in Recommender Systems: A Methodology for Learning, Analyzing, and Mitigating Bias in Ratings. arXiv Link