How to Deradicalize the Recommendation Algorithm: A Statistical Approach

A Data-Driven Framework

José Manuel Rodríguez Caballero

Department of Mathematics and Statistics
Faculty of Science and Engineering
Université Laval

September 21, 2024

Overview

1. Motivation
2. Recommendation Algorithms
3. Hebbian Algorithms
4. Radicalization Pathway
5. Proposed Solution
6. Measuring Success

Hebb's Rule

Neurons that fire together, wire together.

Figure: Hebb's Rule illustrated with memes.

This behavior can lead to correlation between users' preferences over time, potentially spreading both neutral and extreme content through recommendation algorithms.

Recommendation Algorithms

According to NVIDIA[1]:

> *A recommendation system (or recommender system) is a class of machine learning that uses data to help predict, narrow down, and find what people are looking for among an exponentially growing number of options.*

[1] https://www.nvidia.com/en-us/glossary/recommendation-system/

Hebbian Algorithms

A recommendation algorithm is termed *Hebbian* if it satisfies:

> ***Property:*** *If two users (Alice and Bob) share many common products, recommendations for Alice will be influenced by Bob's history, and vice versa.*

Potential Issue: In this framework, users can unknowingly

influence each other's recommendations, leading to unintended sociological consequences.

Example of Hebbian Algorithm

Consider the following simplified model of a Hebbian algorithm:

$$P(\text{Recommend } x \text{ to Alice}) = \frac{\#\text{ of times Bob consumed } x}{\#\text{ of items Bob consumed}}.$$

Comment: This model assumes the influence between users is directly proportional to shared consumption, but real-world systems typically involve more complex patterns.

Pathway to Radicalization

Problem: If Bob is radicalized and Alice shares many common interests, a Hebbian algorithm could unintentionally expose Alice to radical content.

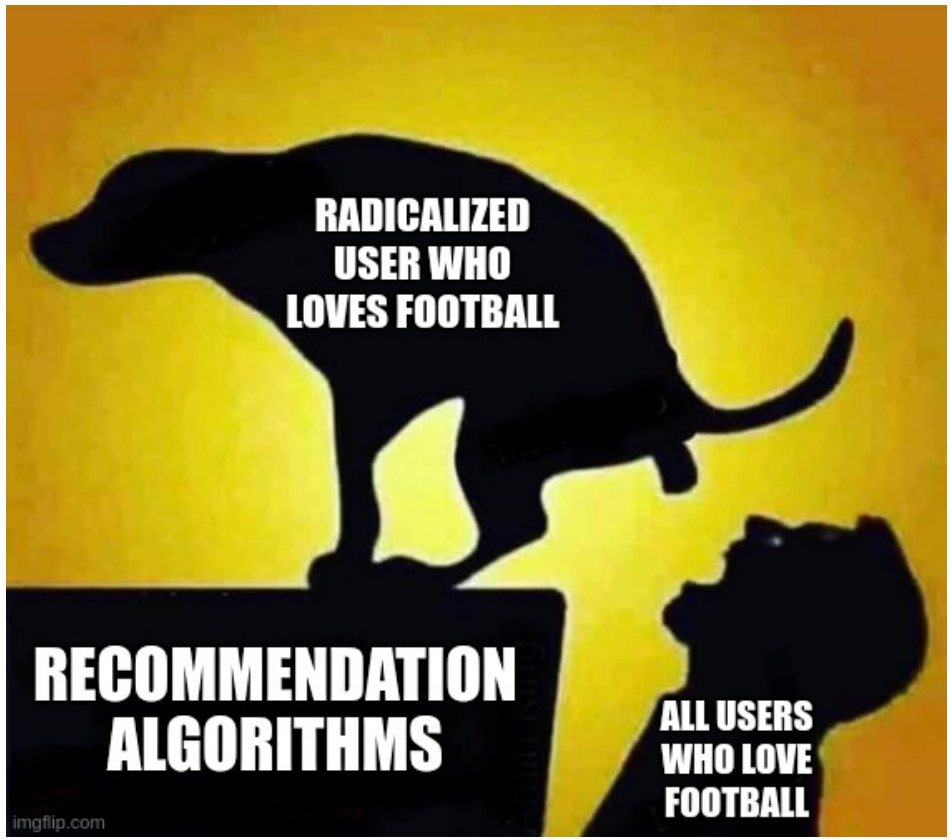

Figure: Radical content spreads via common user connections.

Proposed Solution: Statistical Intervention

1. **Identification of Radicalized Users:**
   - Apply classification algorithms to detect users exhibiting extreme behavior patterns.
2. **Exclusion from Collaborative Filtering:**
   - Ensure that consumption patterns from identified users are excluded from influencing other users' recommendations.
3. **Deradicalization Content:**
   - Target radicalized users with curated content designed to counter extremist views.

Measuring Success: Statistical Approach

Key Evaluation Metric: Reduction in Radicalized Recommendations

1. **Pre-Implementation:**
   - Sample users and record radicalized recommendations over a period (e.g., one month).
2. **Post-Implementation:**
   - Measure radicalized recommendations after applying the exclusion method.
3. **Statistical Testing:**
   - Conduct a *paired t-test* to compare pre- and post-implementation recommendation counts.
   - Estimate the effect size (e.g., Cohen's *d*) to quantify the impact of the intervention.

Thank You!