LessWrong is a community blog and forum about rationality and AI. To preserve its norms amidst a recent influx of newcomers, it has introduced increasingly strict moderation policies, causing some users to complain.

“Growing Pains” example submitted by Sydney A.

What’s with all the bans recently?
by [anonymous]  5 min read  3rd Apr 2024  6 answers  22 comments

Summary: the moderators appear to be soft banning users with 'rate-limits' without feedback. A careful review of each banned user reveals it's common to be banned despite earnestly attempting to contribute to the site. Some of the most intelligent banned users have mainstream instead of EA views on AI.

You can post again in 1 day.
A moderator has rate limited you.

0.5% extra credit for examples relevant to recent or upcoming lectures. Submit on Ed under the “Lectures” category.
Announcements

Project proposal + prototype due next Tuesday

Project proposal will be a contract between you and the TA in terms of the ambition of your project and deployment

Want feedback on your idea? Swing by TA or Michael’s office hours!

Assignment 2 will be released after the project proposal is turned in, and will be due after one week
Most viral memes

Assignment 1: Going Viral

As voted by the class
#1 voted by class
by yabraham
Me to my in-class discussion partner I just spent 20 minutes talking to whose name I no longer remember.

#2 voted by class

by murtazal
Ok who put my bike on the z axis

#3 voted by class
by lwinkley
Never give up. You don't know what life has in store for you.

Viral online (2.2k upvotes)

by Umar P.
Last time: growing pains

Communities can’t maintain the same design as they grow. Newcomers change the dynamics, even if they absorb the norms—and oftentimes they don’t absorb the norms.

Growth begets contention and rulemaking, which can push off newcomers

Moderation and governance are key tools in managing the growth
“In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients.”

- Herb Simon, 1971
Designing for info overload

**Ranking**
Unintuitive mental model, but when it’s right, a feed brings what you want to the front

Facebook  Instagram  TikTok
Twitter  Reddit  TikTok
Pinterest

**Chronological**
Simple mental model but spammy accounts can dominate

Mastodon  iMessage  Discord
Email  WhatsApp
Slack
Designing for info overload

Ranking

Unintuitive mental model, but when it's right, a feed brings what you want to the front

Chronological

Simple mental model but spammy accounts can dominate

How do you think a system should be directing attention in an overloaded community? [1 min]
"Algorithms are unavoidable here. Even sorting posts by friends in chronological order or videos by overall popularity is algorithmic; and often it is unclear there is a single, simple baseline algorithm."

- Dean Eckles, MIT, to the US Senate [2021]
Today

Feed ranking algorithms: how they work

Global rankings a la upvote

Personalized rankings a la the For You Page

Feeds’ objective functions — what are we optimizing for?

Feeds are echo chambers: are feeds echo chambers?

[Gilbert, Bergstrom, and Karahalios 2009]
Global ranking

Used by the traditional Reddit “hot” ranking 🙈 or the Fizz ranking or Hacker News

First shot: how many upvotes does it have?

e.g., 100 upvotes ranked above 10 upvotes

…but this ignores if lots of people saw it but a large % disliked it
Global ranking

Reddit, Fizz, and Hacker News also have downvote data!

Second shot: upvotes – downvotes

e.g., 10 upvotes + 1 downvote ranked above 100 upvotes + 100 downvotes

…but this ranking would never stay fresh! The most popular items of all time would never change.
Global ranking

Final shot: decay over time

\[ \log(\max(\text{upvotes} - \text{downvotes}, 1)) \]

Why take the logarithm?

Because the most popular posts have orders of magnitude more upvotes than others: without the log transform, the top posts would never decay fast enough, relative to the other posts.

Finally, decay the log score over time

(Reddit did a linear penalty, Hacker News is more exponential. The choice depends on what exactly you're aiming for.)
Personalized feed: machine learning

1) Featurize
2) Predict
3) Calculate objective
4) Rank
1) Featurize

- Tie strength w/ MSB: 6
- Content type: mobile phone photo
- Platform: iPhone
- Vision algorithm: stuffed animal, bear
- Text features (e.g., BERT embeddings)
- Interactions so far: 101
- % haha reactions: 15%
- Day of year
- Age of content
- Internet: 10 mbps
Personalized feed: machine learning

1) Featurize
2) Predict
3) Calculate objective
4) Rank
2) Predict

- Tie strength w/ MSB: 6
- Content type: mobile phone photo
- Platform: iPhone
- Vision algorithm: stuffed animal, bear
- Text features (e.g., BERT embeddings)
- Interactions so far: 101
- % haha reactions: 15%
- Day of year
- Age of content
- Internet: 10 mbps
2) Predict

Probability of clicking “like”
P(like)
2) Predict

Models for each outcome of interest

- P(like)
- P(hide)
- P(comment)
- P(follow)
- P(watch)
- P(share)
- P(click)
Prediction signals used at Instagram

[https://transparency.fb.com/features/explaining-ranking/ig-feed]

How likely you are to click on the profile of a post’s author

Signals influencing this prediction include:

- How many times you’ve viewed the author’s profile and posts from that author
- How many times people have clicked on the profile of the author
- How many times the author’s profile has been clicked by followers

How likely you are to share content through an external platform by using the "Copy link" or "Share" options

Signals influencing this prediction include:

- How many times you’ve viewed posts from the author
- How much time you’ve spent viewing posts from the author
- Where data privacy laws permit, how many times you’ve shared a post externally, and how many times this post has been shared externally by others

How likely you are to spend little time viewing a post and not interact with it

Signals influencing this prediction include:

- How much time you’ve spent viewing posts from the author
- How many times you’ve viewed at least one post either within your feed or anywhere on Instagram
- The number of posts available for ranking

How likely you are to scroll to the next post

Signals influencing this prediction include:

- How many times you’ve viewed or scrolled to the next post
- How many times others have scrolled to the next post after viewing this post

How likely you are to click a reel to view it in full screen

Signals influencing this prediction include:

- How many reels you’ve clicked to view in full screen
- The total number of times the post has been clicked to be viewed in full screen when displayed in the top position

How likely you are to comment on a post

Signals influencing this prediction include:

- How many times you’ve liked the author’s posts in your feed
- The number of comments the post received overall
- How many comments have been made on posts by the author, in total and by you

How likely you are to comment on a post

Signals influencing this prediction include:

- How many times you’ve clicked to comment on posts
- How many times you’ve clicked “View all comments” on posts

How likely you are to report a post

Signals influencing this prediction include:

- How many times the author recovered their account after it was hacked
2) Predict

How do we train these deep learning algorithms?

Training data: prior behavior on the platform

As you browse, scroll, and click, and as others do, the system builds models to predict your behavior toward future unseen posts.
Personalized feed: machine learning

1) Featurize
2) Predict
3) Calculate objective
4) Rank
3) Calculate objective

So what do we do with all of these predictions?

P(like) P(watch) P(comment) P(click)

P(hide) P(follow) P(share)

We define an objective: an algorithm to combine and weight the predictions

Intuitively: how many points does each predicted behavior get?

\[ \sum_{p \in \text{predictions}} \text{weight}_p \cdot p \]
Objectives in use

Facebook: “Meaningful Social Interactions”: a weighted average of Likes, Reactions, Reshares, and Comments

<table>
<thead>
<tr>
<th>Interaction type</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like</td>
<td>1</td>
</tr>
<tr>
<td>Reaction</td>
<td>1.5</td>
</tr>
<tr>
<td>Reshare</td>
<td>1.5</td>
</tr>
<tr>
<td>Comment</td>
<td>15-20</td>
</tr>
</tbody>
</table>

Source: https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Facebook overhauls News Feed in favor of 'meaningful social interactions'

Refresh of the News Feed algorithm will de-prioritize content shared by media and businesses in favor of that produced by friends and family, Zuckerberg said.

Julia Carrie Wong in San Francisco
Objectives in use

Twitter’s open sourced algorithm:

- 75 points if predicted that, if you reply, the author will reply back
- 27 points if predicted that you’d reply
- 12 points if predicted that you engage with the author’s Twitter profile
- 1 point if predicted that you’ll retweet
- 0.5 points if predicted that you’ll favorite
- -74 points if predicted that you’ll give negative feedback (“not interested”, mute, block)
- -369 points if predicted that you’ll report it
Objectives in use

TikTok: watching, liking, commenting

The document says watch time isn’t the only factor TikTok considers. The document offers a rough equation for how videos are scored, in which a prediction driven by machine learning and actual user behavior are summed up for each of three bits of data: likes, comments and playtime, as well as an indication that the video has been played:

\[ \sum_{p \in \text{predictions}} \text{weight}_p \cdot p \]
This is why we talk about feeds as being driven by engagement

Engagement is typically a shorthand for behaviors that the platform can observe: e.g., likes

But, optimizing for engagement can create negative outcomes. What are they, and what could we do about them? [2min]
This is also why feeds include predictions for global objectives

**Indirect impacts**: if we show this to you, and you leave a comment, will it make a better experience for the user who posted it?

**Long-term impacts**: what impact will this have on your wellbeing? [Burke and Kraut 2016; Stray 2020]
Impacts estimated via survey: machine learning models trained on survey responses — some users have surveys injected into their feed where they rate whether a particular item is important, informative, funny, or makes them feel connected [Eckles 2021]

Feed diversity: penalize feeds that are all the same kind of content
Sum it up

\[ \sum_{p \in \text{predictions}} \text{weight}_p \cdot p \]

Content

P(like) P(watch) P(comment) P(click) P(share) P(follow) P(hide)

Score: 2.4
Sum it up

\[
\sum_{p \in \text{predictions}} \text{weight}_p \cdot p
\]

Score: 1.1
Personalized feed: machine learning

1) Featurize
2) Predict
3) Calculate objective
4) Rank
4) Rank

Rank the items in the feed by their score.
Outcomes of ranking

Ranking (on Twitter)

- See content from more accounts
- More political content

Chronological (on Twitter)

- A few “loud” accounts dominate
- More external links

[Bandy and Diakopoulos 2021]
How decisions get made

Typically, the platform runs an A/B test on, say, 1% of its users to test the impact of a feed ranking change on its metrics.

Often, in practice, the criterion is, “does this move up your desired metrics without harming our other metrics?”

So the team trained a machine-learning algorithm to predict posts that users would consider “bad for the world” and demote them in news feeds. In early tests, the new algorithm successfully reduced the visibility of objectionable content. But it also lowered the number of times users opened Facebook, an internal metric known as “sessions” that executives monitor closely.

“The results were good except that it led to a decrease in sessions, which motivated us to try a different approach,” according to a summary of the results, which was posted to Facebook’s internal network and reviewed by The Times.
Topics I won’t cover today

“Why is TikTok’s feed so good?”

It’s not the algorithm — it’s the signals

Inventory: can it choose from every piece of content on the platform, or just the accounts you follow?

Embeddings: by structuring the deep learning model as a recommender system (think: Netflix), it can jumpstart its recommendations.
Feeds are echo chambers: are feeds echo chambers?

[Gilbert, Bergstrom, and Karahalios 2009]
Filter bubbles

Filter bubbles occur when everyone is shown only content that they like. This happens as a natural outcome of optimizing for engagement.

Example: YouTube recommendation radicalization: channels that are slightly less mainstream become recommendation gateways to more and more radical channels [Ribeiro et al. 2020]
Echo chambers

If feed algorithms only show you things that you want to see, and only shows me things that I want to see, then…

Won’t the end result be an echo chamber, where we only hear people who share our opinions? Won’t this further polarize our society?
We can look at Facebook log data to understand where polarization might be coming from [Bakshy, Messing, and Adamic 2015; González-Bailón et al. 2023]

The biggest drop is due to homophily, not the algorithm: your post inventory is typically filled with like-minded friends. It drops further as we go down the funnel from inventory to exposure to engagement.
Those who use social media are exposed to more cross-cutting ideological news than those who don’t use social media [Fletcher and Nielsen 2017]

Subscribing people to counter-partisan news sources in their feeds decreases negative attitudes toward the other political party by only 1 point on a 100-point scale [Levy 2021; Bail et al. 2018]

Simulations suggest that we might have it backward: that it’s not that we’re polarized because social media only exposes us to like minds, but we’re polarized because social media exposes us to a wider variety of people [Törnberg 2022]
An experiment during election season switched over 20,000 Facebook and Instagram users from a ranked feed to a reverse chronological feed [Guess et al. 2023]

The chronological feed substantially reduced how much time people spent on the platform, and exposed people to more political content.

But, it didn’t impact political polarization.
Summary

One common strategy for managing growth is to decide on a subset of content to show users, through an algorithmic feed:

- Global rankings aggregate up/downvotes, then trail off over time.
- Personalized rankings predict on-platform behaviors, then assign weights to each predicted behavior to determine a score.

Concerns abound about feeds creating filter bubbles and echo chambers. While there are clearly negative outcomes, the science is now catching up to what turns out to be a complicated story.
References


References


Social Computing
CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.

Slide content shareable under a Creative Commons Attribution-NonCommercial 4.0 International License.