Understanding and Improving Content Moderation Systems in Web3 Platforms

Wenrui Zuo
Queen Mary University of London

Aravindh Raman
Telefónica Research

Raul J Mondragón
Queen Mary University of London

Gareth Tyson
Hong Kong University of Science & Technology (GZ)

* Published in Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2024)

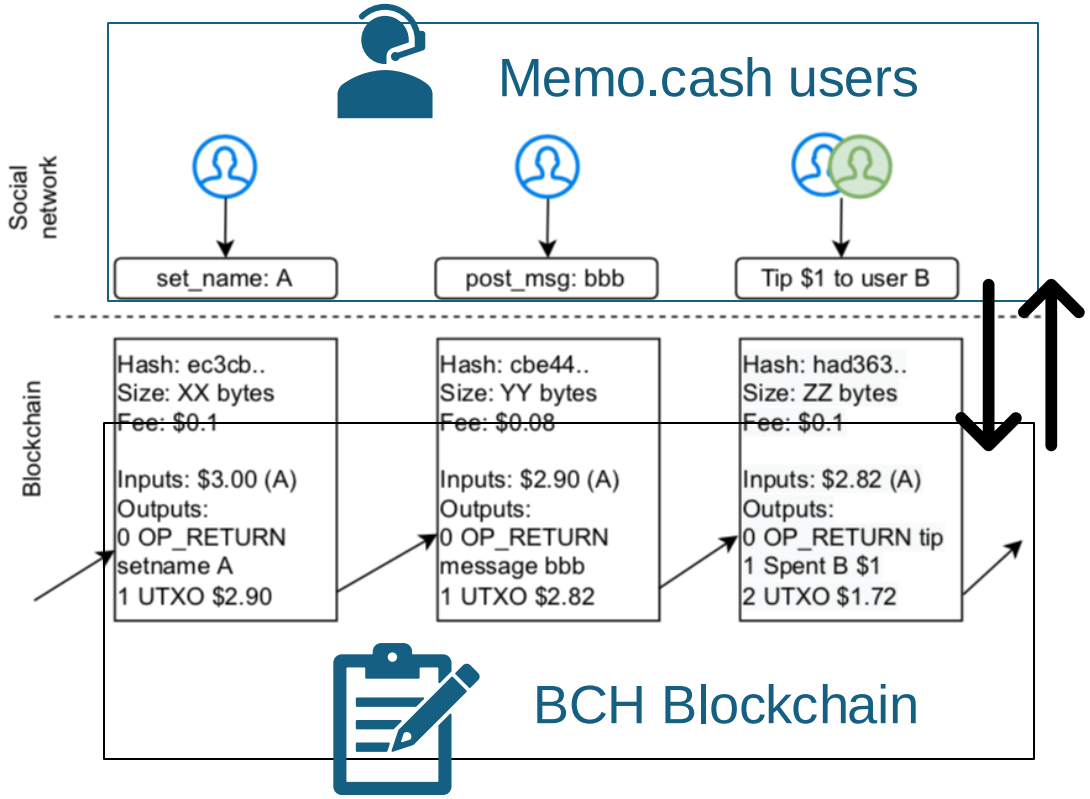
Recap

- Content Moderation on Web3 Social Media
- The Reason Behind Mutes
- Mute Localization
- Mute-list Recommendation

Content Moderation on Web3 Social Media (memo.cash)

- memo.cash is social media built on top of Bitcoin Cash Blockchain
  - ***Permanent*** and uncensorable data
  - ***User-controlled*** moderation
    - Individuals create publicly available mute-list
    - Similar to blocking on Twitter

- Dataset
  - Develop a ***crawler*** to parse web pages
    - 24K users
    - 317K posts
    - 2M transactions
    - 7K mutes



The Reason Behind Mutes

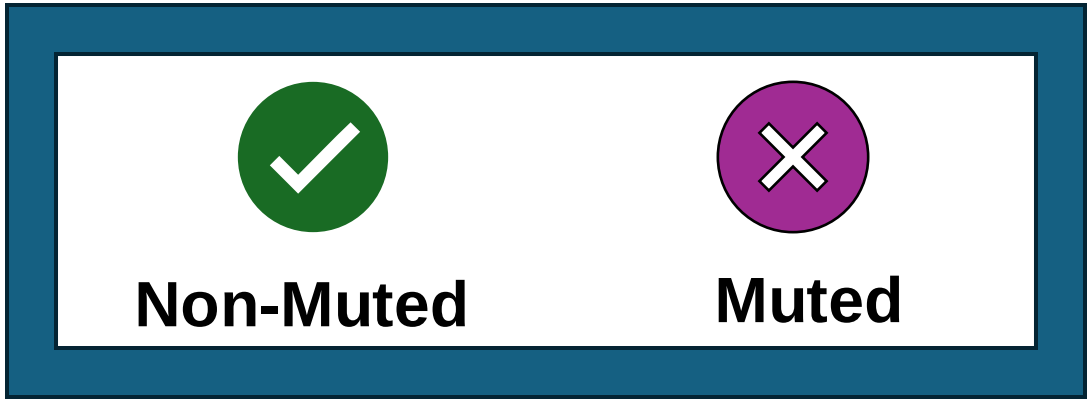
User-based Information

Bitcoin Transaction

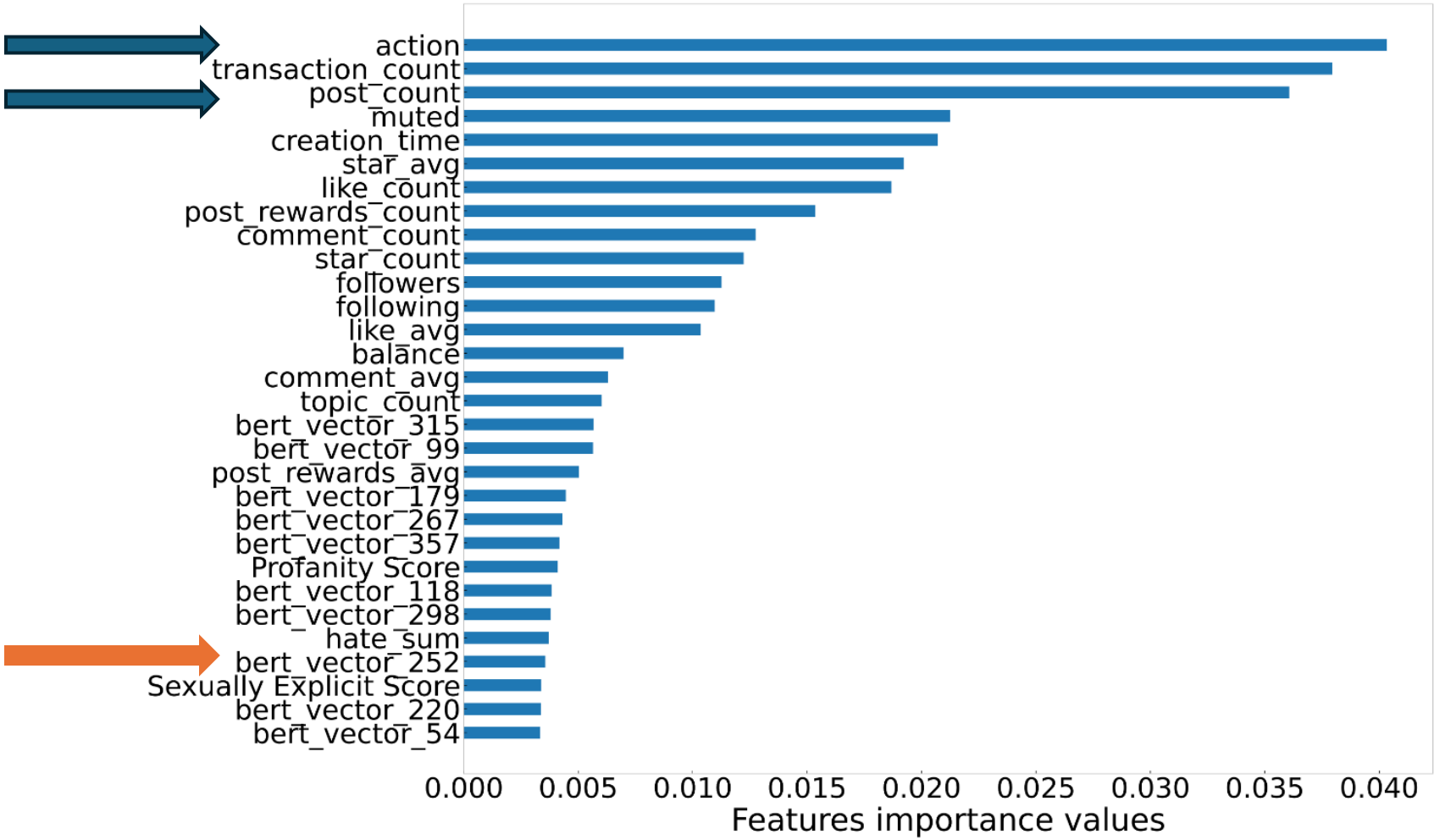
Posted Content

Toxic/Hateful Speech

Social Interaction

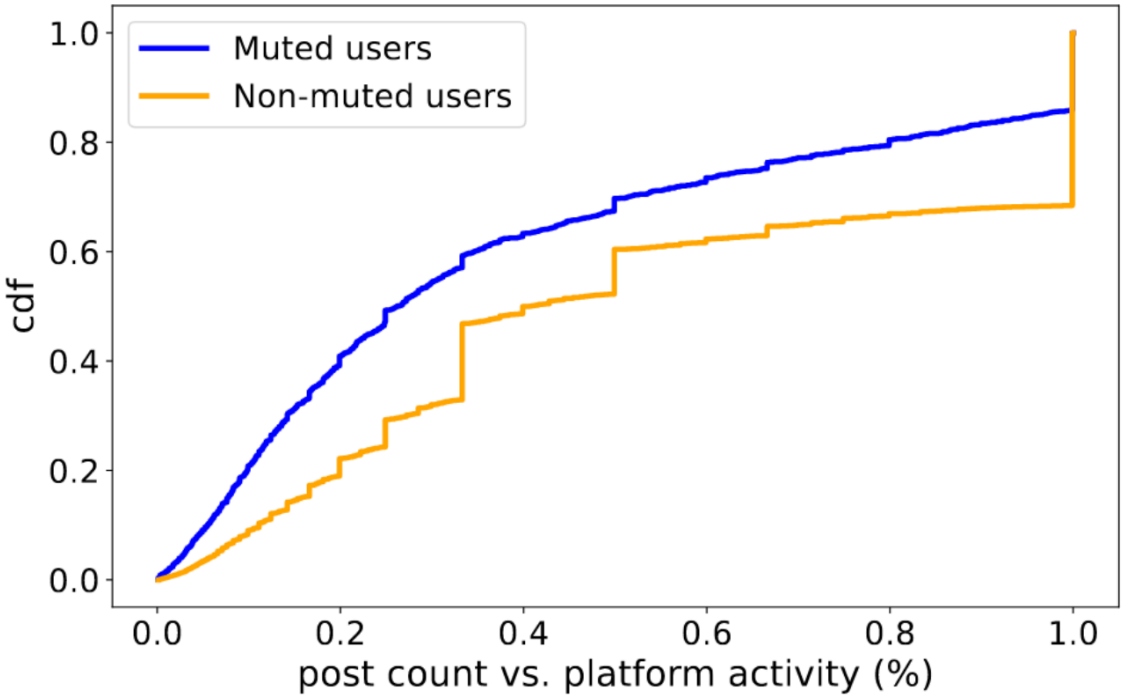


The Reason Behind Mutes



Hateful speech emerges as a less significant factor, while users' activeness is the most important!

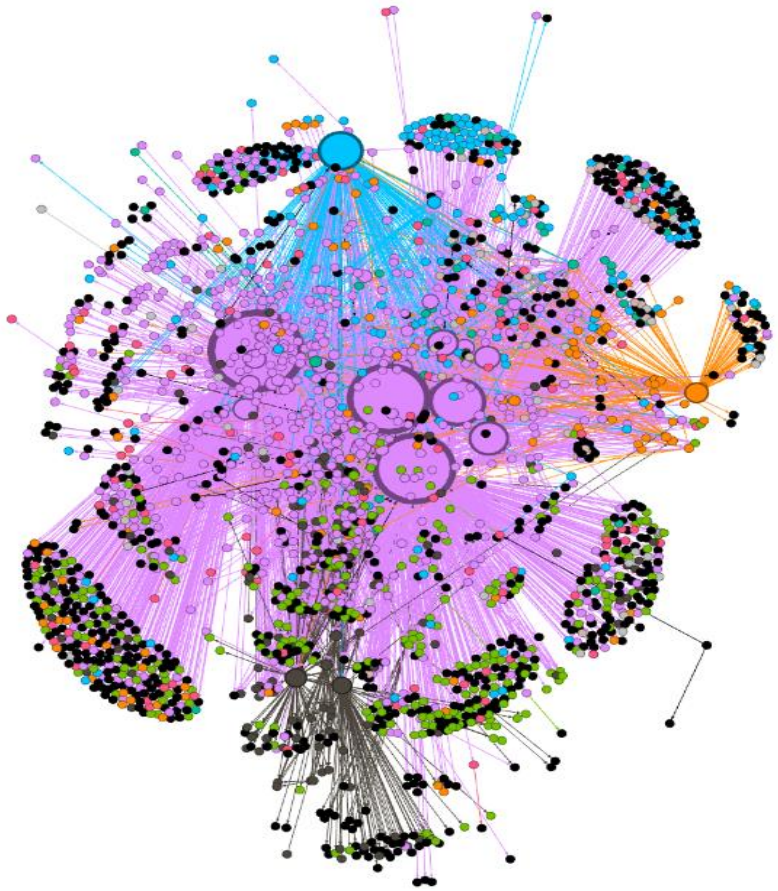
The Reason Behind Mutes



- 19.6% of muted users (vs. 33.1% non-muted) have a ratio > 0.8
  - Suggesting higher average platform engagement among muted users
  - Muted users have a time interval of 215s (vs. non-muted 437s)
  - 60.5% of muted (vs. 47.0% non-muted) users' posts get 0 likes, tips, or replies

The presence of low-quality or irrelevant content could be a contributing factor prompting users to resort to muting

Mute Localization



Mute graph

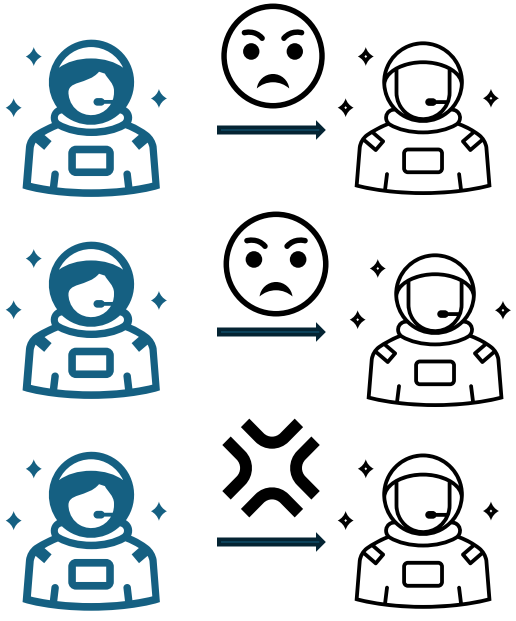
- The followership network comprises 12,676 nodes and 60,809 links
  - 11 main communities (Louvain Method)
- Visualization of the mute graph
  - Node colour is based on the corresponding community on the followership network
  - Node size is determined by mute count

Mute Localization

- Massey Denton Isolation Index to define the mute localization across communities
- Compare Real isolation index (RI) value and random simulation index (SI)
  - 9 out of 11 communities have RI > SI values
  - 60.9% of the mutes originate from users belonging to the same community

Users are more inclined to mute others in the same community!

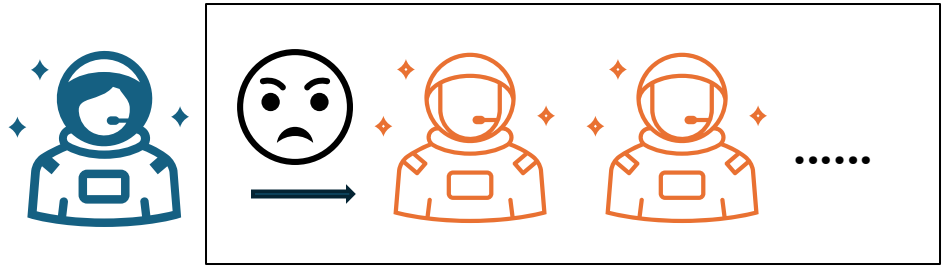
Mute-list Recommendation



Input

Random
Most Popular
SVD
UserKNN
LightFM
CNNbaseline
CNN-preprocess
CNN-normalised
CNNpreprocess+
normalised

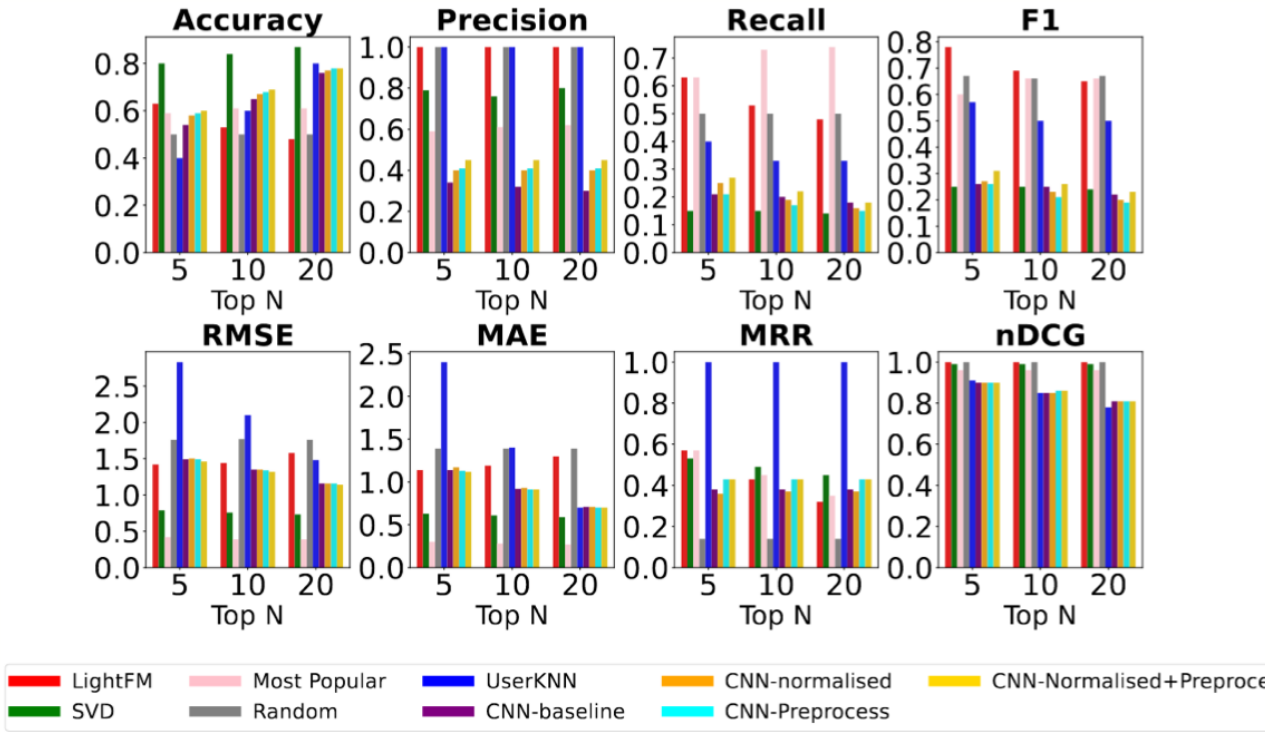
Output

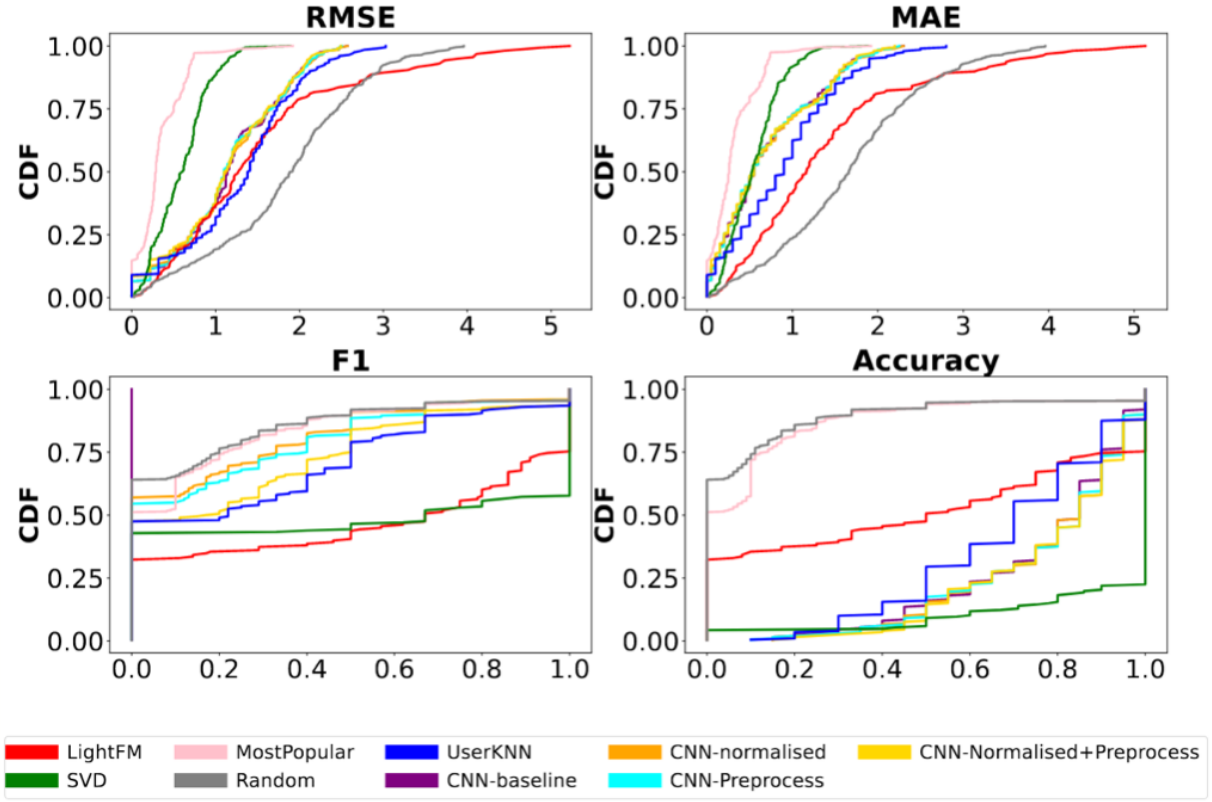


Personalized top-n mute list

Label	Implication	Event Count
4	Very negative feedback: User A has muted user B, and user A's follower has done the same. Additionally, user A has not liked any user B posts.	525
3	Moderately negative feedback: While user A has muted user B, their followers have not taken such action. Nevertheless, user A has never liked user B's posts.	1,825
2	Somewhat negative feedback: User A has muted user B, and their follower has also done the same. However, user A has liked at least one of user B's posts.	896
1	Slightly negative feedback: Despite user A muting user B, user A has also liked a post by user B.	2,996

Mute-list Recommendation

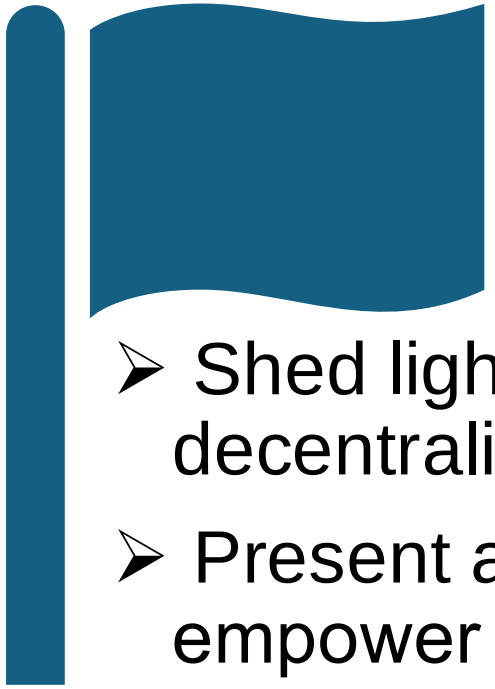




- LightFM exhibits strong performance across all metrics, particularly in top-10 mute list scenario

- LightFM is the most suitable model
- 38.8% of users exhibit F1 values <0.5
- 41.4% of users display accuracy values <0.5

5. Conclusion



- Shed light on the ***key user's factors*** impacting Web3 decentralized moderation
- Present a novel approach for recommending mutes to empower users' participants

Thanks for Listening!