Content Moderation

Antriksh Ganjoo
6th May 2024

Agenda

1. Dive Deeper into content moderation
2. Overview of the Chapters
3. Facts and Figures related to content moderation

What is Content Moderation?

Why do we need Content Moderation?

1. Social Media Content Moderation acts as a filter to remove harmful content of posts and to make it safe for the targeted audience.
2. Unfiltered content has a high potential to play havoc of all sorts and thus raise the need to moderate these posts.

Chapter 4

- Three imperfect solutions to the problem of scale

3 Imperfect Solutions:

- Editorial Review
- Community Flagging
- Automatic Detection

Editorial Review overview:

1. Editorial review refers to the process by which platforms employ human moderators to review and make decisions about user-generated content, often based on predefined community standards or guidelines.

2. Gillespie highlights the role of human moderators in interpreting and applying platform policies in nuanced and context-dependent ways.

3. Unlike automated moderation systems, which rely on algorithms to flag and remove content based on predefined criteria, editorial review involves human judgment and discretion in assessing the appropriateness of content.

Case Study: Apple

1. App store is the only place where users can get IOS apps.
2. SDKs are also highly moderated by Apple
3. Specific apps are allowed to be published on the App store

Discussion 1

1. What role should governments play in regulating content moderation practices on social media platforms, if any?

Community Flagging:

- Community flagging, also known as user-generated content moderation, involves allowing users themselves to report or flag content they deem inappropriate or violating platform guidelines.

- Community flagging plays a crucial role in social media content moderation, fostering a collaborative approach to maintaining a safe and inclusive online environment.

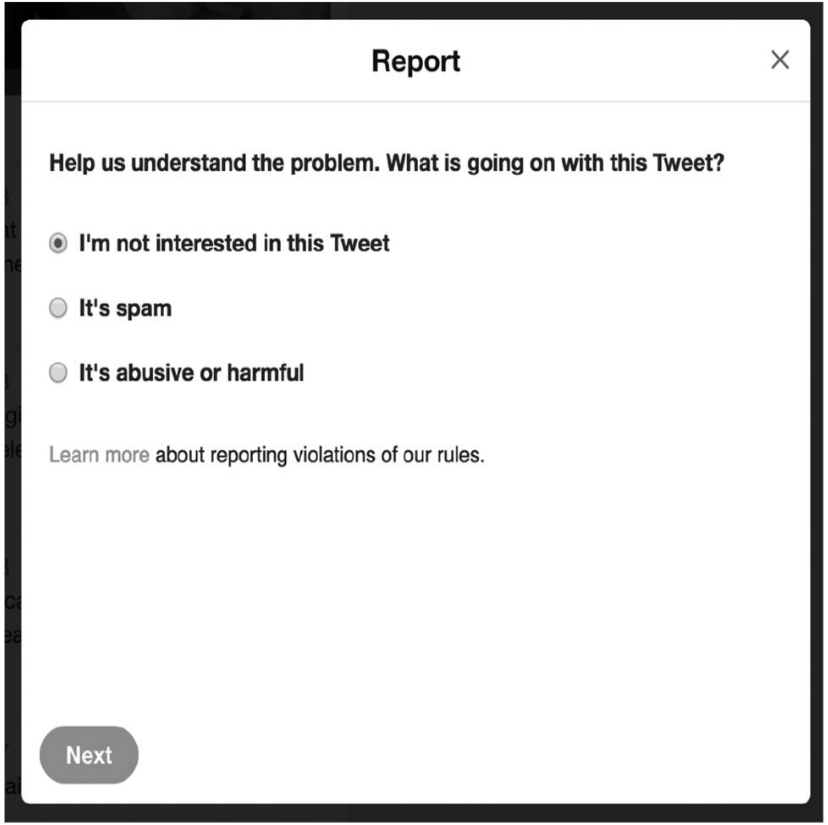
- However, it requires careful management and oversight to address challenges such as abuse and bias effectively.

Flagging in Social Media Platforms:

- Enhances User Participation: Flagging empowers users to contribute to the moderation process by identifying problematic content.

- Scalability: Helps platforms manage the vast volume of user-generated content by prioritizing moderation efforts based on flagged content.

- Community Standards: Supports the enforcement of platform policies and community guidelines by highlighting content that violates established rules.



Report ✕

Help us understand the problem. What is going on with this Tweet?

● I'm not interested in this Tweet

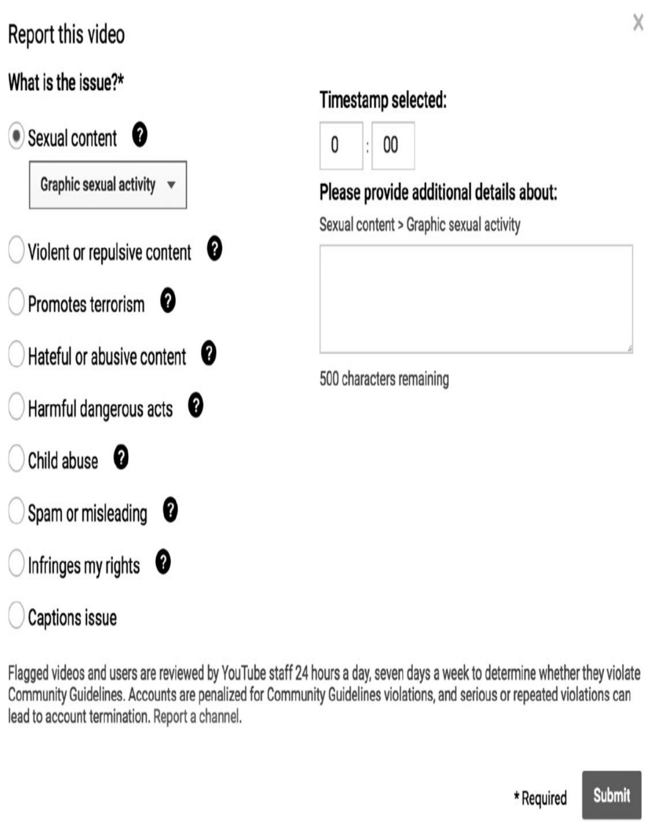
○ It's spam

○ It's abusive or harmful

Learn more about reporting violations of our rules.

Next

Twitter, flagging pop-up window (2017)



Report this video X

What is the issue?*

● Sexual content

Graphic sexual activity ▾

○ Violent or repulsive content

○ Promotes terrorism

○ Hateful or abusive content

○ Harmful dangerous acts

○ Child abuse

○ Spam or misleading

○ Infringes my rights

○ Captions issue

Timestamp selected:

0 : 00

Please provide additional details about:

Sexual content > Graphic sexual activity

500 characters remaining

Flagged videos and users are reviewed by YouTube staff 24 hours a day, seven days a week to determine whether they violate Community Guidelines. Accounts are penalized for Community Guidelines violations, and serious or repeated violations can lead to account termination. Report a channel.

*Required Submit

YouTube, flagging pop-up window (2017)

Discussion 2

2. What strategies can platforms implement to increase transparency and accountability in their content moderation processes?

Is Flagging a complete solution?

- False Positives: Flagging can result in the removal of content that does not actually violate platform guidelines, leading to unintended censorship.

- Abuse: Users may misuse the flagging system to target content they disagree with or to harass other users.

- Limited Context: Flagging often relies on users' subjective interpretations of content, which may lack context or nuance.

Discussion 3

3. Should users have more control over the content moderation settings on social media platforms, and if so, how could this be implemented?

Automatic Detection:

- ❖ Multimedia Analysis: AI can analyze not only text-based content but also images, videos, and audio files, expanding the scope of content moderation to multimedia formats.

- ❖ Consistency: AI algorithms can apply content moderation rules consistently across all user-generated content, reducing the risk of bias or subjectivity inherent in human moderation decisions.

- ❖ Continuous Improvement: AI algorithms can learn and adapt over time based on feedback and new data, continuously improving their accuracy and effectiveness in identifying and moderating content violations.

Discussion 4

4. Is AI the answer to content moderation ?

Chapter 6

Facebook, breastfeeding and living with suspension

Case Study: Breastfeeding

- Gillespie examines how Facebook's initial policy classified breastfeeding images as a violation of its community standards on nudity and sexual content, leading to the suspension of users who shared such images.
- This policy sparked significant controversy and criticism from users who argued that breastfeeding is a natural and important act that should not be censored.
- The conclusion drawn from the case of Facebook, breastfeeding, and suspension is that content moderation is a complex and evolving process influenced by a range of factors, including cultural norms, community standards, user feedback, and platform policies.
- Gillespie emphasizes the importance of ongoing dialogue and engagement between platforms and users to address the challenges and controversies surrounding content moderation effectively.

Discussion 5

3. What do you think is the effect of removing these images on the users who post these images on their channel/page?

Chapter 8

What Platforms are, What they should be

Ways to Improve Moderation:

- ➔ Putting Real Diversity Behind the Platform
  - ◆ Empathy and Understanding: Moderators from diverse backgrounds are often more empathetic and understanding towards the experiences and challenges faced by marginalized or underrepresented groups.
  - ◆ Cultural Competence: Moderators from diverse backgrounds bring cultural competence and awareness to the moderation process, enabling them to better understand and interpret content within its cultural context.

Ways to Improve Moderation:

- Transparent Guidelines and Policies:
  - Platforms should provide clear and accessible guidelines and policies outlining their content moderation rules and standards.
  - This transparency enables users to understand what constitutes acceptable behavior on the platform and what types of content may be subject to moderation.
- Algorithmic Transparency:
  - Platforms should strive to make their content moderation algorithms more transparent and understandable to users.
  - While some aspects of AI-powered moderation may be proprietary, platforms can still provide insights into how algorithms operate and the factors considered in moderation decisions.

Ways to Improve Moderation:

➔ Reject the economics of popularity

- Mitigating Harmful Content: Popular content is not always synonymous with responsible or constructive content. By moving away from popularity-driven algorithms, platforms can reduce the risk of harmful or misleading content being amplified simply because it generates high engagement.

- User Empowerment: Rejecting the economics of popularity empowers users to curate their own content experiences based on their interests, values, and preferences, rather than being solely at the mercy of algorithmic recommendations driven by popularity metrics.

Discussion 6

6. Can content moderation ever be completely unbiased, or is it inherently influenced by the values and perspectives of platform owners and moderators?

Some Facts and Figures related to Content Moderation

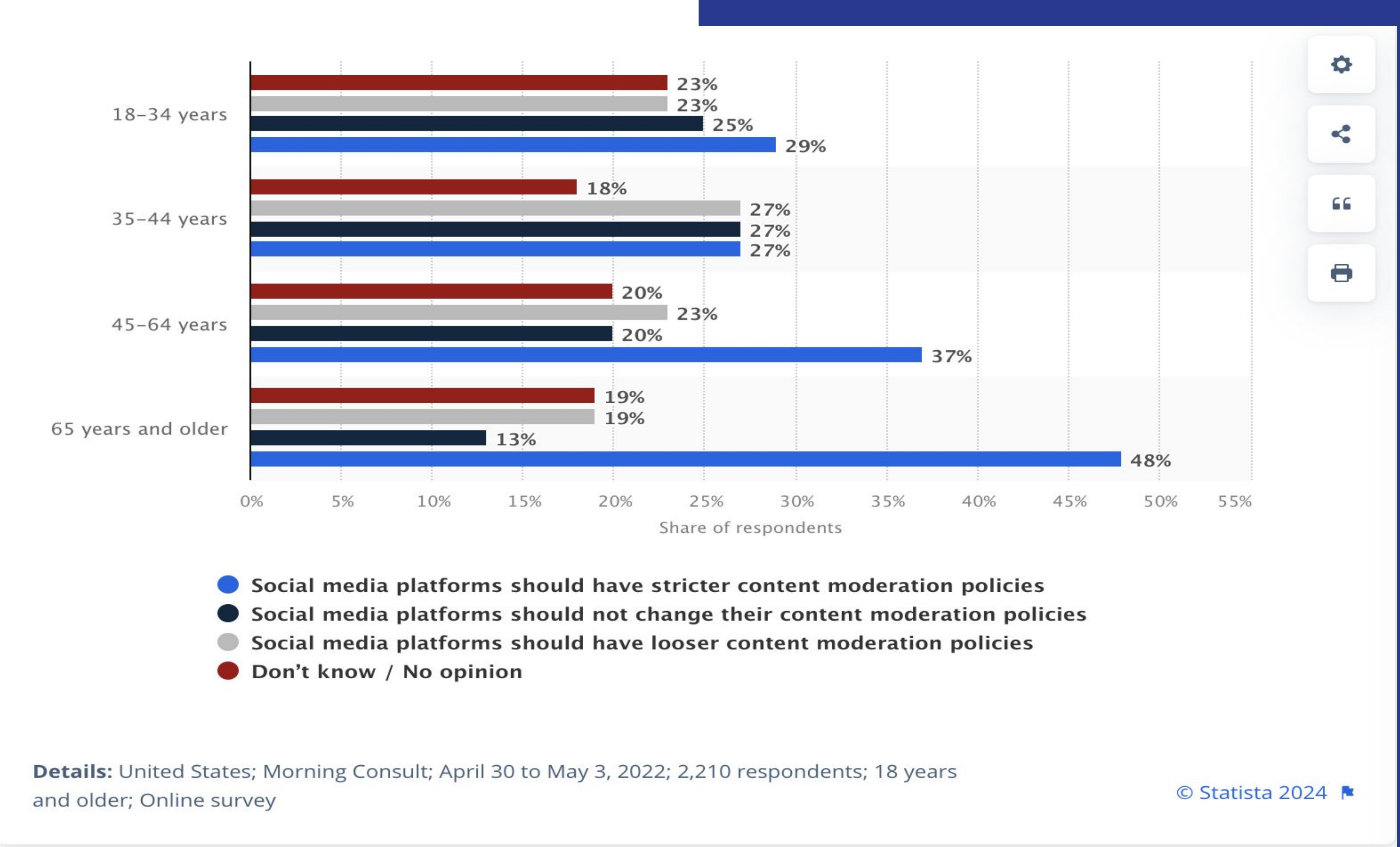
The rate of moderation automation is very high at Meta: At Facebook and Instagram, respectively, 94% and 98% of decisions are made by machines – far more than the 45% reported by TikTok.

6 BILLION POSTS were removed in the second half of 2020.Social media sites are clearly working hard to removehuge swathes of harmful content online.

41% of American adults have experienced online harassment, and 66% of adults have witnessed at least one harassing behavior online. – Pew Research

81 percent of women and 43 percent of men had experienced some form of sexual harassment during their lifetime. – GfK

And the list goes on ....



Details: United States; Morning Consult; April 30 to May 3, 2022; 2,210 respondents; 18 years and older; Online survey

© Statista 2024

Accounts Suspended for Community Guideline Violations
(July-December 2020)

	Twitter	TikTok	YouTube	Pinterest	Snap
Accounts Suspended	1,009,083	6,144,040	3,859,685	4,269	47,558

Number of Posts Removed by Rule Violation
(July-December 2020)

	Twitter	Facebook	Instagram	TikTok	YouTube	Pinterest	Snap
Graphic Violence	59,933	34,800,000	9,700,000	267,398	13,861	1,754	337,710
Child Sexual Exploitation/ Minor Safety	9,178	17,800,000	1,809,400	32,087,857	40,383	1,794	47,550 (accounts deleted for CSAM specifically)
Hateful Content	1,628,281	49,000,000	13,100,000	1,782,658	42,013	2,487	77,587
Abuse Or Harassment	1,448,418	9,800,000	7,600,000	5,882,773	79,902	3,763	238,997

References & Citations:

- https://www.lemonde.fr/en/pixels/article/2023/11/14/content-moderation-key-facts-to-learn-from-facebook-instagram-x-and-tiktok-transparency-reports_6252988_13.html#:~:text=The%20rate%20of%20moderation%20automation,the%2045%25%20reported%20by%20TikTok.
- https://netchoice.org/wp-content/uploads/2021/11/Content-Moderation-By-The-Numbers-v5.pdf
- https://www.statista.com/statistics/1361043/us-adults-opinion-social-media-content-moderation-by-age/
- https://www.statista.com/topics/11495/social-media-content-moderation-and-removal/#topicOverview

Any questions or ideas to further discuss?

THANK YOU!