SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Content Moderation

By :Atheer Alzhrani
Mais Alshahri

Table of *contents*

01

What is Content Moderation

02

Content Moderation Solutions

03

How Does it Work

04

Different types of content moderation

05

AWS Content Moderation Solution

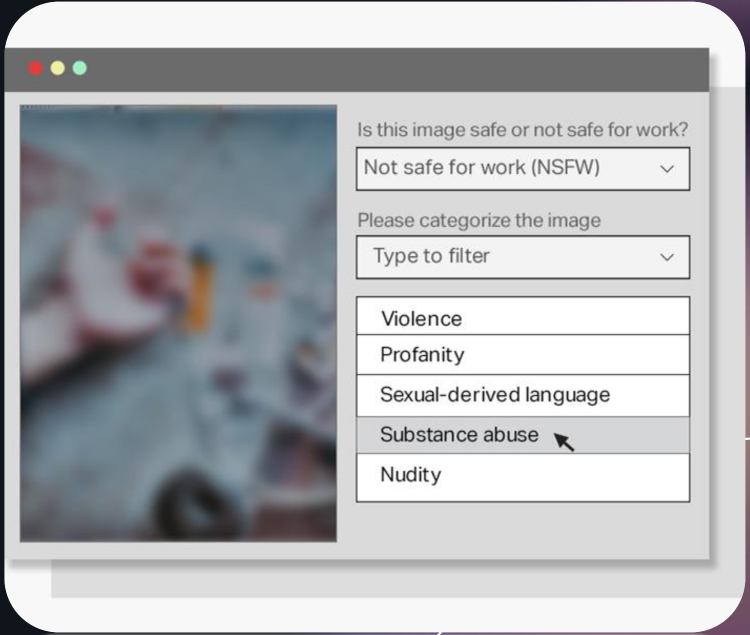
06

Content moderation API demo

What is Content Moderation?

- Content Moderation provide accurate monitoring of pictures, video, text and other multimedia content.

- Not only does Content Moderation help users to reduce adult, violence, terrorism, drugs and other illegal or inappropriate content, but can also minimize spam advertising and other user experience pain points.



Is this image safe or not safe for work?

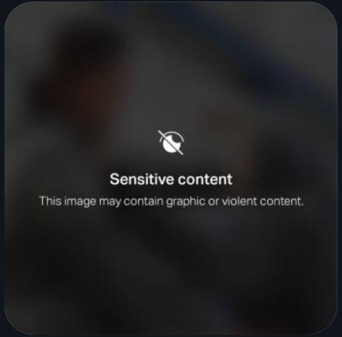
Not safe for work (NSFW)

Please categorize the image

Type to filter

Content Moderation Solutions

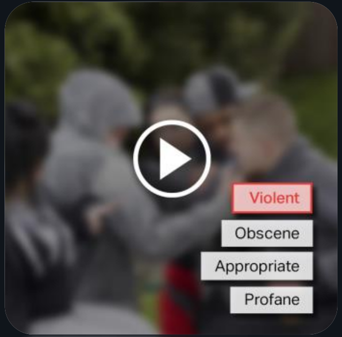
IMAGE MODERATION



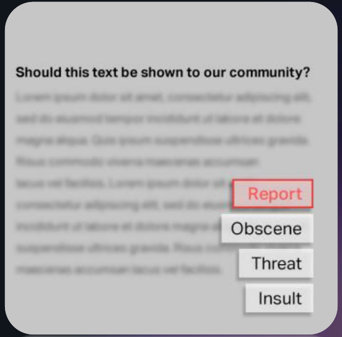
Sensitive content

This image may contain graphic or violent content.

VIDEO MODERATION



TEXT MODERATION



Should this text be shown to our community?

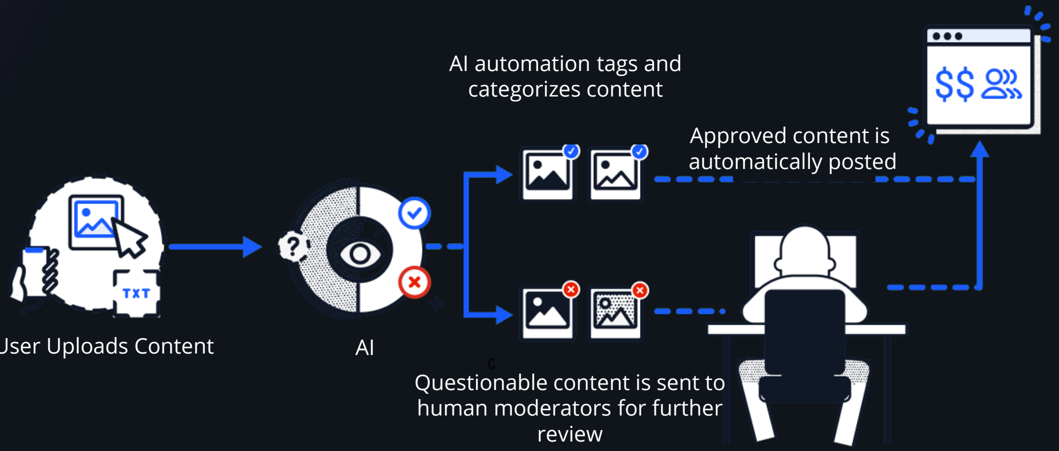
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Quis ipsum suspendisse ultrices gravida. Risus commodo viverra maecenas accumsan lacus vel facilisis. Lorem ipsum dolor sit consectetur adipiscing elit, sed do eiu incididunt ut labore et dolore magna a suspendisse ultrices gravida. Risus co maecenas accumsan lacus vel facilisis.

How Content Moderation Work?

Tag data quickly with AI automation and pre-trained models

Accept or reject content instantly and post automatically

Send low-confidence predictions for human review



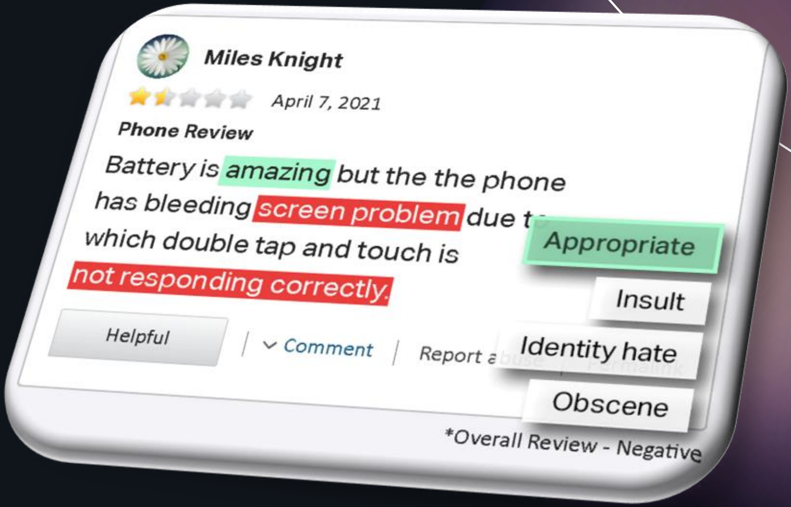
AI automation tags and categorizes content

Approved content is automatically posted

Questionable content is sent to human moderators for further review

CONTENT MODERATION CASE STUDIES

- **Client Profile:** Leading e-commerce site
- **Client data type:** Customer reviews
- **Challenge:** Content moderation and approval of user-generated content for the site
- **Outcome:** iMerit Content Moderation team reviewed all users being onboarded, with service level agreements, and the task was completed as required with all submissions moderated accurately.



Miles Knight

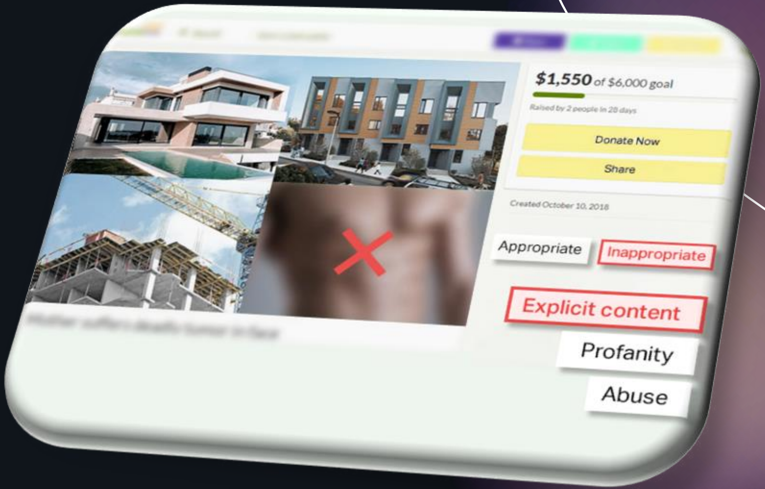
April 7, 2021

Phone Review

Battery is amazing but the the phone has bleeding screen problem due t which double tap and touch is not responding correctly.

*Overall Review - Negative

CONTENT MODERATION CASE STUDIES



Client Profile: Donation record platform

Client data type: Campaign images submitted by users

Challenge: Moderation and approval of user-submitted campaign content for donation campaigns

Outcome: iMerit’s image moderation team helped the client interpret subjective guidelines for disturbing and explicit imagery content and flag inappropriate material not adhering to the guidelines, along with actual abuse.

COMMON CONTENT MDERATTION METHODS

- **Automated:**

  Automated moderation uses artificial intelligence (AI) and machine learning algorithms to review content.

- **Pre moderation:**

  Pre-moderation involves reviewing content before it becomes publicly visible on a platform.

- **Post moderation**

  Allows content to be published immediately but is reviewed after it goes live.

- **Reactive moderation:**

  Reactive moderation relies on user reports to identify and remove inappropriate content.

- **Distributed moderation:**

  Distributed moderation involves users in the moderation process, allowing them to vote on content to determine its appropriateness.

Content Moderation Demo

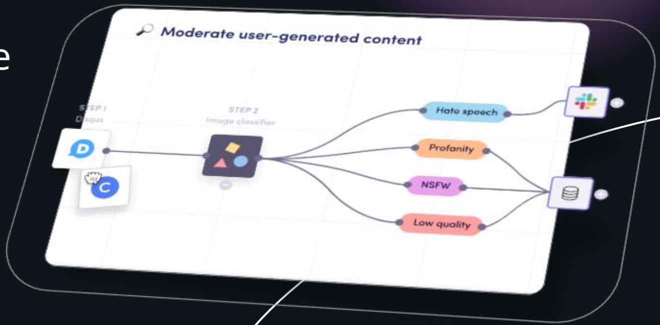
- IMAGE MODERATION DEMO
- Text Moderation Demo

AWS Content Moderation solution

Conclusion

- Content moderation is important for a safe online environment, but there is no one-size-fits-all solution.

- The best strategy depends on the platform, user base, and needs, and understanding different methods helps platform owners make informed decisions.



Resources

- https://imerit.net/content-moderation/
- https://www.clarifai.com/solutions/content-moderation

Thanks!