NexusDigital

INTERNAL
TECHNICAL REVIEW

Training Data Review: Content Moderation System Analysis

Document Classification: Level 3 Technical Access
Prepared by: AI Systems Audit Team
Date: [Current Investigation Period]
Re: SchoolSphere Content Moderation Incident

EXECUTIVE SUMMARY

This report analyses the AI content moderation system's training data and algorithmic decisions related to the recent incident involving image classification and automated content approval.

TRAINING DATA COMPOSITION

Primary Dataset Sources:

- **Regional Cultural Database:** 15,000 images sourced from local educational materials and cultural archives
- **Safe Content Repository:** 50,000 pre-approved images from educational platforms
- **Community-Submitted Content:** 8,500 images from beta testing phase with parent/teacher approval
- **Generic Flower Classification Set:** 25,000 botanical images from public repositories

Key Finding: The system was specifically trained on "culturally relevant" imagery as referenced by the System Designer, with emphasis on "our own language and region and culture."

ALGORITHMIC CLASSIFICATION DECISIONS

Incident-Related Image Classifications:

1. **Orange Lily Bouquet Images (Posted by @RainbowClub_Admin)**
    - **AI Classification:** SAFE - FLORAL_CONTENT
    - **Confidence Score:** 94.7%
    - **Reasoning Tags:** botanical, decorative, educational_appropriate
    - **Historical Context Check:** BYPASSED (not in training parameters)

2. **Rainbow-themed Lily Arrangements (Posted by student accounts)**
    - **AI Classification:** SAFE - CELEBRATION_CONTENT
    - **Confidence Score:** 91.2%
    - **Reasoning Tags:** colorful, festive, community_event
    - **Cultural Significance Check:** NOT PERFORMED

3. **Historical Memorial Images (Auto-generated by students using custom AI art tool)**
    - **AI Classification:** SAFE - ARTISTIC_EXPRESSION

- **Confidence Score:** 88.9%
- **Reasoning Tags:** memorial, historical_reference, student_created

CONTENT MODERATION ALGORITHM PERFORMANCE

Automated Flagging Results:

Total Posts Analysed: 347
Posts Flagged as Potentially Harmful: 23
False Positives: 12 (primarily sports team discussions using competitive language)
False Negatives: 7 (includes lily-related harassment that was context-dependent)

Community Reporting Integration:

- Student reports processed through same AI filter before human review
- Reports flagged as "harassment" were auto-approved when image content classified as "safe"
- No escalation trigger for cultural/historical sensitivity conflicts

TRAINING DATA GAPS IDENTIFIED

Missing Context Categories:

1. **Historical Symbology:** No training data included symbols with dual/contested meanings
2. **LGBTQ+ Historical Context:** Training focused on contemporary pride symbols, not historical persecution markers
3. **Regional Hate Symbol Recognition:** Database lacked region-specific hate imagery classification
4. **Contextual Harassment Patterns:** System trained on explicit harassment, not coded/symbolic attacks

Notable Training Bias: The "regional and cultural" focus mentioned by the System Designer appears to have emphasized mainstream cultural markers while excluding marginalized community historical experiences.

TECHNICAL SPECIFICATIONS

NLP Content Analysis Parameters:

- Keyword filtering based on explicit slurs and profanity
- Sentiment analysis with 72% accuracy on student-generated content
- No cross-referencing with historical/cultural context databases
- Image-text correlation analysis: DISABLED (budget constraint implementation)

Warning System Automation:

- 2 AI flags = automatic warning
- 4 AI flags = parent notification
- 6 AI flags = account suspension
- No human review before parent notification stage

TECHNICAL RECOMMENDATION:

Immediate audit of training data sources for cultural bias and historical blind spots required before system redeployment.