Color Guard - Research

0 Backgrounds

1. Main Issues in NLP Detecting Harmful Content
    a. A lack of data in low-resource settings
    b. Inconsistent definitions and criteria for judging harmful content, requiring classification models to be robust to spurious features and diverse.
2. ...

1 NLP tools

1. Detoxify
    - https://github.com/unitaryai/detoxify?utm_source=chatgpt.com
    - Developed by Unitary, Detoxify comprises trained models capable of predicting toxic comments across multiple datasets, including those from Jigsaw's challenges. It utilizes deep learning techniques to classify content into categories like threats, obscenity, insults, and identity-based hate.
2. ToxiCraft
    - https://aclanthology.org/2024.findings-emnlp.970/?utm_source=chatgpt.com

- This framework focuses on synthesizing datasets of harmful information using a small amount of seed data. ToxiCraft generates a wide variety of synthetic yet realistic examples of toxic information, enhancing the robustness and adaptability of detection models.

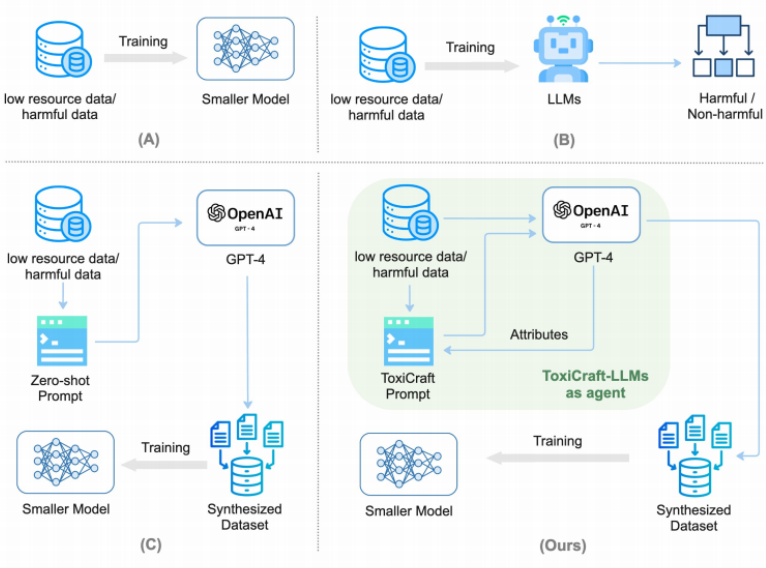

Figure 1: Harmful detection approaches

-

3. BD-LLM (Bootstrapping and Distilling Large Language Models)
   - BD-LLM employs a novel prompting method called Decision-Tree-of-Thought (DToT) to improve large language models' performance in toxic content detection. It enhances detection efficiency and extracts high-quality rationales for better interpretability.
4. Lettria's NLP Platform
   - Lettria leverages machine learning and neural networks to detect and moderate harmful content. Their platform analyzes text data to identify abusive language and can take actions such as hiding or removing the content. It also detects users who frequently post harmful content, allowing for preventive measures.
5. Supporting Human Raters with Large Language Models
   - This approach explores leveraging large language models to assist human raters in identifying harmful content, including hate speech, harassment, and misinformation. It proposes design patterns that integrate LLMs with human rating to enhance detection accuracy and efficiency.
6. Perplexity-based Detection Methods
   - These methods utilize perplexity measures to identify adult and harmful content in multilingual, heterogeneous web data. By training models on harmful textual data, they can effectively distinguish between harmful and non-harmful content based on perplexity thresholds.

2 Framework

2.1 Data Resources

Public Posts

- Photos
- Comments
- Tags

```
Technical Considerations:
    Ensure user consent for data collection, anonymization,
    and legal compliance with data privacy laws.
Ethical Considerations:
    Respect users' privacy and only use publicly available information.
```

Previous Violations

- Past reports or bans

Engagements

- User interactions (likes, shares, messages)

2.2 Behavior Assessment

2.2.1 NLP Model Role

- Analyze text content to detect harmful language and behaviors using NLP tools.
- Utilize advanced NLP models like Detoxify or ToxiCraft to ensure robust detection.
- * Ensure fairness in NLP model training and prevent biases in detection algorithms.

2.2.2 Behavior Assessment Criteria

1) NLP Detection in Behavior Scoring

a. **Weighted Contribution**:

- NLP-detected behaviors (e.g., harmful language, bullying) will be assigned a predefined weight in the overall behavior assessment score.
- For example:
  - **60%**: NLP detection results for harmful language or inappropriate content.

- **40%**: Historical behavior (e.g., previous reports, frequency of flagged content).

b. **Severity Levels**:

i. NLP results will classify harmful behaviors into severity levels (e.g., minor, moderate, severe), influencing the score differently:

1. **Minor**: Small deduction in score.
2. **Moderate**: Moderate deduction.
3. **Severe**: Significant deduction.

c. **Behavior Categories**:

- NLP will map detected behaviors to specific categories (e.g., hate speech, grooming, threats).
- Each category will have a unique impact factor, ensuring nuanced scoring.

d. **Cumulative Scoring**:

- The system will maintain a cumulative scoring model where repeated or escalating harmful behaviors, as detected by NLP, progressively reduce the user’ s behavior score.

e. **Time Decay**:

- Scores can recover over time if no further harmful behaviors are detected, promoting improvement and fair evaluation.

2) Previous Violations

a. **Historical Score Deductions**:

- Each past ban or violation is assigned a penalty that directly affects the assessment score.
- Penalty severity is determined based on:
  - **Frequency**: Number of past violations.
  - **Severity**: Nature of the violations (e.g., temporary suspension vs. permanent ban).

b. **Penalty Decay Over Time**:

- Older bans or violations have diminishing impact over time to allow for behavior improvement.
- Example:
  - Violations within the last 6 months: **Full weight** in penalty.
  - Violations older than a year: **Reduced weight** or ignored.

c. **Violation Categories**:

- Classify past bans into categories (e.g., bullying, hate speech, spam).
- Assign specific weightings to these categories to reflect their impact on the safety score.

d. **Cumulative History Factor**:
   - Incorporate a **history factor** into the overall score formula:
   - $BehaviorScore = BaseScore - (NLPDeductions + HistoryPenalty)$
   - Example:
     - Base Score: 100
     - NLP Deductions: 15
     - History Penalty: 10 (based on two past bans)
     - Final Score: 100−(15+10)=75

e. **Special Cases for Persistent Violators**:
   - Users with recurring violations may receive compounded penalties or additional scrutiny, lowering their score more significantly.

Symbolic Scoring System

$$BehaviorScore = BaseScore - (\alpha * NLPDeductions + \beta * HistoryPenalty)$$

- $BaseScore$：100
- α： 0.6 （NLP detection results for harmful language or inappropriate content）
- β： 0.4 （Historical behavior）
- NLP Deductions:

Equation:

$$S = \sum_{i=1}^{n} (W_s \cdot C_i \cdot R_i \cdot D)$$

- S: NLP deduction score (final output)
- Ws: Severity weight for each behavior detected (Ws=w1,w2,w3, representing minor, moderate, severe)
- Ci: Impact factor for specific behavior category i (e.g., hate speech, grooming, threats)
- Ri: Frequency of detected behavior for category i (how many times the behavior has occurred)

- D: Time decay factor (reduces the impact of older harmful behaviors)

Explanation of Terms:

1. **Severity Weight (Ws)**:
   - Minor: Ws=0.5
   - Moderate: Ws=1.0
   - Severe: Ws=2.0
2. **Impact Factor (Ci)**:
   - Hate speech: Ci=1.5
   - Grooming: Ci=2.0
   - Threats: Ci=2.5
3. **Frequency (Ri)**:
   - For each behavior category, count the number of times the behavior was detected.
4. **Time Decay (D)**:
   - Recent events: D=1.0
   - 6–12 months old: D=0.75
   - Over 1 year old: D=0.5

Example Calculation:

- **Detected Behaviors**:
  - 3 instances of minor hate speech (detected recently)
  - 2 instances of severe grooming (detected 8 months ago)
- **Using the Equation**:
  - $S=(0.5 * 1.5 * 3 * 1.0)+(2.0 * 2.0 * 2 * 0.75)$
  - $S=(2.25) + (6.0)$
  - $S=8.25$

2.2.3 Assessment Result

- Provide a score (0-100) reflecting the user's vehavior quality, based on NLP analysis and other inputs.
- Implement scalable scoring logic and real-time updates for dynamic behavioral changes.

- Avoid stigmatization of users with low scores; promote improvement instead.

2.3 Color Coding System

2.3.1 Base Color Bar

The **color bar** visually represents a scale from 0 to 100, where each point is assigned a unique color gradient transitioning smoothly from red to green.

- **Leftmost (0 points)**: Bright red, indicating aggressive or highly inappropriate behavior.
- **Rightmost (100 points)**: Bright green, indicating safe and appropriate behavior.
- **Midpoints (e.g., 50 points)**: Transition through colors like orange and yellow, signaling cautionary zones.

2.3.2 How the Color Bar Works:

1. **Color Gradient**:
   - The system generates a continuous gradient of colors using RGB or HSL values.
   - For example:
     - 0 points: RGB(255, 0, 0) (red).
     - 50 points: RGB(255, 255, 0) (yellow).
     - 100 points: RGB(0, 255, 0) (green).
2. **Behavior Assessment Score Mapping**:
   - Each user’ s behavior score (0–100) determines their position on the color bar.
   - Scores closer to 0 are flagged as “high risk” (red zone).
   - Scores closer to 100 are flagged as “safe” (green zone).
3. **Dynamic Updates**:
   - User scores are recalculated in real time based on recent behavior, ensuring the color reflects their current status.

2.3.3 System Functionality Breakdown

1. **NLP Detection**:
   - NLP models analyze user content and assign deductions based on the type, severity, and frequency of harmful behavior.
   - Outputs include:
     - Behavior categories (e.g., hate speech, bullying).
     - Severity levels (minor, moderate, severe).

- A cumulative deduction score.

2. **Behavior Assessment**:
   - Combines NLP deductions, historical penalties, and time-decay factors to calculate a behavior score.
   - The score formula: Behavior Score=100−(NLP Deductions+Historical Penalties)\text{Behavior Score} = 100 - (\text{NLP Deductions} + \text{Historical Penalties})Behavior Score=100−(NLP Deductions+Historical Penalties)
3. **Color Assignment**:
   - The behavior score is mapped to a corresponding color on the color bar.
   - For example:
     - 0–20: Red (high risk).
     - 21–50: Orange (moderate risk).
     - 51–80: Yellow (low risk).
     - 81–100: Green (safe).
4. **Visibility for Minors**:
   - The color-coded score is only visible to minors, alerting them to potential risks.
   - It appears as a small icon or label next to the user’ s name/profile.

2.3.4 User Interface Example

1. **Color Bar Display**:
   - A horizontal gradient bar, ranging from red to green, is used to visually explain the scoring system.
   - Each user is shown their assigned score and color.
2. **Icons and Notifications**:
   - Red Zone: Displays warnings like “This user has been flagged for inappropriate behavior.”
   - Green Zone: Displays “This user is safe for interaction.”
3. **Transparency for Adults**:
   - Adults are notified if their behavior affects their score, with recommendations for improvement.

2.4 User Safety Mechanism

- Marks and color codes are visible only to minors to guide their interaction decisions.
  - Prevent misuse or overexposure of marks, maintaining user trust and ethical handling.
  - Protect minors' sensitive data and identities while displaying adult behavior ratings.
- Alerts minors about potentially harmful interactions, enabling interactions, enabling them to avoid engaging with flagged users.
  - Develop safeguards against false positives/negatives and provide user appeal mechanisms.
  - Promote education on cyber safety and respectful communication.