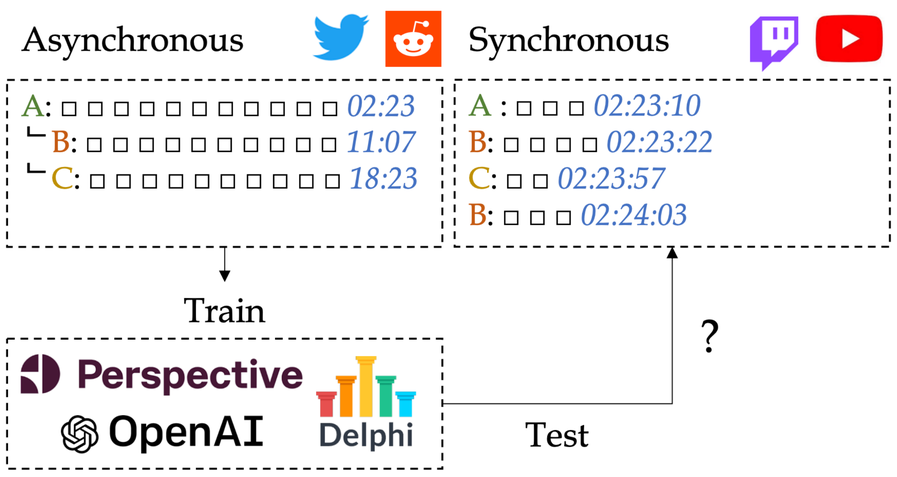
# Analyzing Norm Violations in Real-Time Live-Streaming Chat

Jihyung Moon*[1], **Dong-Ho Lee***[1,2], Hyundong Cho[2], Woojeong Jin[2], Chan Young Park[3], Minwoo Kim[4], Jonathan May[2], Jay Pujara[2] and Sungjoon Park[1]

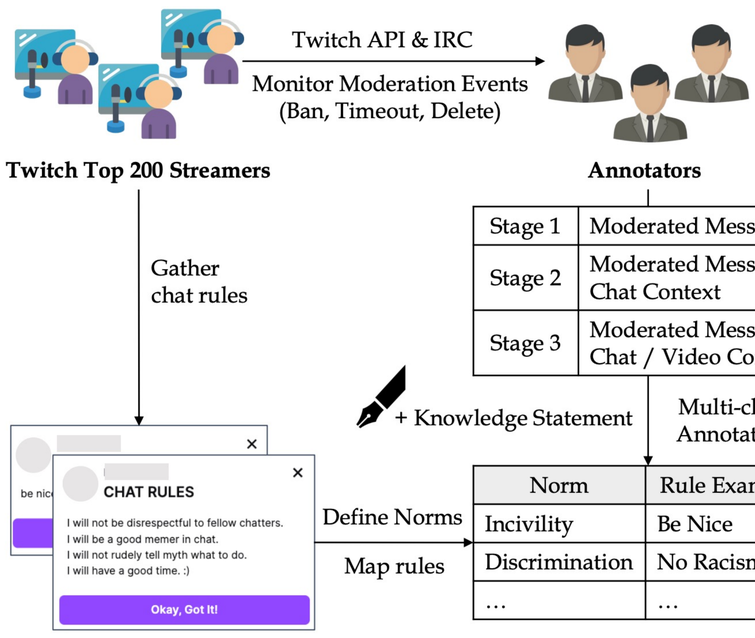
[1]SoftlyAI Research, [2] University of Southern California, [3]Carnegie Mellon University, [4] Selectstar

## Motivation




Asynchronous domain has …
(1) Smaller message length and temporal gap between chats;
(2) Relationships between chats are less clearly defined.

## Data Construction



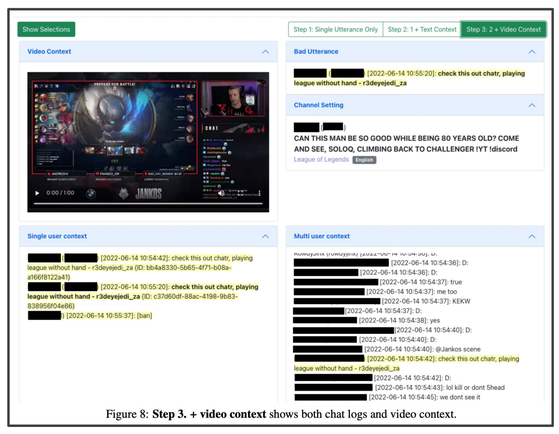




Figure 8: **Step 3. + video context** shows both chat logs and video context.

Annotation Tool

## Norms

| Coarse | Fine-grained | Rule Examples |
|---|---|---|
| Discrimination | Discrimination | No racism, sexism or homophobia. |
| HIB (Harassment, Intimidation, Bullying) | HIB | No HIB towards broadcaster<br>No HIB towards viewers, moderators, etc.<br>No HIB towards other broadcasters, politicians, etc. |
| Privacy | Doxing | Please no personal questions about me.<br>Do not share any personal information about yourself or others. |
| Inappropriate Contents | NSFW<br>Self-destructive<br>Illegal<br>Spoiler | No NSFW content (e.g., Inappropriate ASCII arts).<br>No talk of suicide.<br>No drug discussion of any kind.<br>Do not give game spoilers. |
| Off Topic | Controversial Topic<br>Begging | No drama, politics or religion.<br>No begging for subscriptions or money. |
| Spam | Excessive & Repetitive<br>Advertisements | No walls of text.<br>No self promotion unless authorized. |
| Meta-Rules (Live streaming specific) | Backseating & Tall order<br>Mentioning other broadcasters<br>Specific language only | Don't tell me what to do. Don't ask for mod.<br>Don't talk down on other streamers.<br>English only. |
| Incivility (Miscellaneous) | Incivility | Be nice, Be civil |

**Iterative coding process**: authors individually code for rule types with certain categories, come together to determine differences, and then repeat that coding process.

Create categories based on 329 chat rules written by 200 Twitch streamers.

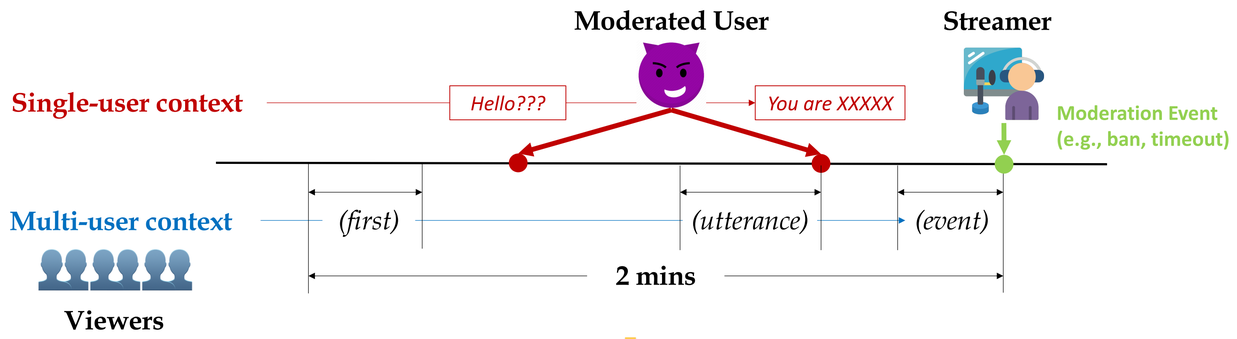
## Data Statistics

| Coarse | Fine-grained | # Rules | # Violates stage 1 | # Violates stage 2 | # Violates stage 3 |
|---|---|---|---|---|---|
| Discrimination | Discrimination | 13.98% (46) | 2.34% (104) | 2.25% (101) | 2.34% (105) |
| HIB | HIB | 22.49% (74) | 21.33% (947) | 26.55% (1,190) | 27.80% (1,246) |
| Privacy | Doxing | 0.60% (2) | 0.34% (15) | 0.36% (16) | 0.36% (16) |
| Inappropriate Contents | Spoiler | 0.60% (2) | 0.02% (1) | 0.02% (1) | 0.02% (1) |
| | NSFW | 1.82% (6) | 0.86% (38) | 0.85% (38) | 0.85% (38) |
| | Self-destructive | 1.21% (4) | 0.32% (14) | 0.29% (13) | 0.29% (13) |
| | Illegal | 0.30% (1) | 0.16% (7) | 0.07% (3) | 0.07% (3) |
| Off Topic | Controversial Topic | 5.47% (18) | 0.59% (26) | 0.85% (38) | 0.83% (37) |
| | Begging | 1.51% (5) | 1.44% (64) | 1.36% (61) | 1.36% (61) |
| Spam | Excessive & Repetitive | 11.24% (37) | 17.59% (781) | 21.64% (970) | 21.42% (960) |
| | Advertisements | 11.24% (37) | 4.64% (206) | 4.40% (197) | 4.42% (198) |
| Meta-Rules (Live streaming specific) | Mentioning other streamers | 14.28% (47) | 0.72% (32) | 10.62% (476) | 10.58% (474) |
| | Backseating & Tall order | 5.16% (17) | 3.45% (153) | 3.70% (166) | 3.77% (169) |
| | Specific language only | 10.03% (33) | 0.97% (43) | 6.94% (311) | 6.94% (311) |
| Incivility (Miscellaneous) | Incivility | - | 12.30% (546) | 11.57% (519) | 11.51% (516) |
| | Non-Identifiable | - | 32.93% (1,462) | 8.52% (382) | 7.45% (334) |
| Total | | 329 | 4,439 | 4,482 | 4,482 |

**Similarity** between Reddit (*Synchronous*) & Twitch (*Asynchronous*):
(1) Harassment (Discrimination, HIB) and Incivility take up a large portion.
(2) Off-topic, Inappropriate contents, Privacy exist but less enforced.

**Difference** between Reddit (*Synchronous*) & Twitch (*Asynchronous*):
(1) Spam, Meta-Rules cover significantly higher portion in Twitch.
(2) Fewer rules about contents. → Streamers are less concerned about contents.

## Different contexts




## Performance of Existing Models

| Model | Precision | Recall | F1 |
|---|---|---|---|
| ToxiGen | 0.31 | 0.91 | 0.46 |
| Perspective API | 0.39 | 0.95 | 0.56 |
| OpenAI moderation | 0.11 | 0.94 | 0.20 |
| OpenAI content filter | 0.55 | 0.86 | 0.67 |

Table 5: **Performance (Binary F1) of toxicity detection models** on HIB and Discrimination data. Binary F1 refers to the results for the 'toxic' class.

Existing models do not frequently produce false positives **(high recall)**,

but perform poorly in detecting toxic messages, with a detection rate of 55% **(low precision)**

## Performance on Norm Classification

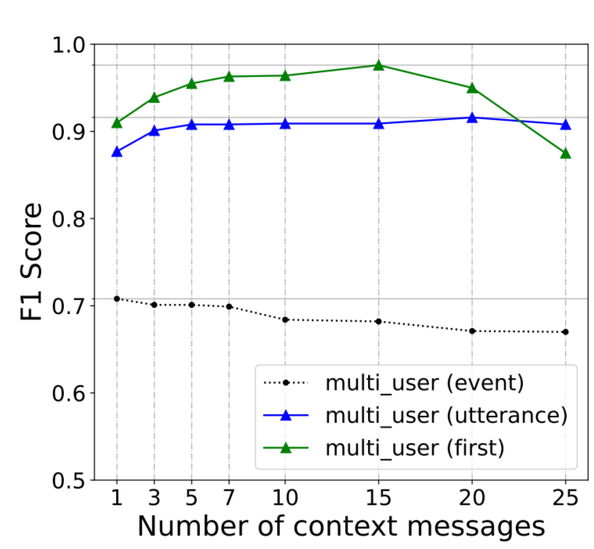
| Context | All | HIB | Off Topic | Spam | Meta-Rules | Incivility |
|---|---|---|---|---|---|---|
| - | 0.70 | 0.52 | 0.07 | 0.63 | **0.65** | 0.28 |
| Single-user context | 0.78 | 0.50 | 0.05 | 0.67 | 0.58 | 0.28 |
| Multi-user context (event) | 0.75 | 0.44 | 0.05 | 0.66 | 0.60 | 0.17 |
| Multi-user context (utterance) | 0.91 | **0.61** | 0.10 | 0.66 | **0.65** | 0.24 |
| Multi-user context (first) | **0.95** | **0.61** | 0.08 | **0.70** | 0.62 | **0.45** |
| Broadcast category | 0.77 | 0.48 | 0.13 | 0.65 | 0.64 | 0.30 |
| Rule text | 0.75 | 0.11 | **0.29** | 0.58 | 0.38 | 0.13 |

**Multi-user context:** Chats from other users help determine the toxicity

**Multi-user context (first):** Temporal gap between event and the actual offending chat may be substantial

Overall, **Context** matters a lot for training better norm classification model.

## Context Size




15~20 messages help the most.

## Distribution Shift between Reddit and Twitch

| | Train data → | Reddit | Twitch | Twitch | Reddit |
|---|---|---|---|---|---|
| Reddit (Normvio) | Twitch (Normvio-RT) | R | T→R | T | R→T |
| ALL | ALL | 0.99 | 0.98 | 0.91 | 0.67 |
| Incivility | Incivility | 0.74 | 0.56 | 0.24 | 0.09 |
| Harassment | HIB, Privacy | 0.41 | 0.20 | 0.62 | 0.26 |
| Spam | Spam | 0.53 | 0.27 | 0.66 | 0.28 |
| Off Topic | Off Topic | 0.28 | 0.12 | 0.10 | 0.00 |
| Hate Speech | Discrimination | 0.19 | 0.06 | 0.04 | 0.02 |
| Content | Inapt. Contents | 0.37 | 0.05 | 0.09 | 0.00 |
| | Test data → | Reddit | | Twitch | |

Models trained on Twitch **generalize better** than models trained on Reddit **despite having 6X less training data**

https://github.com/softly-ai/live-NormVio