EMNLP 2023

Analyzing Norm Violations in Real-Time Live-Streaming Chat

Jihyung Moon*, **Dong-Ho Lee***, Hyundong Cho, Woojeong Jin, Chan Young Park,
Minwoo Kim, Jonathan May, Jay Pujara, Sungjoon Park

softly.ai

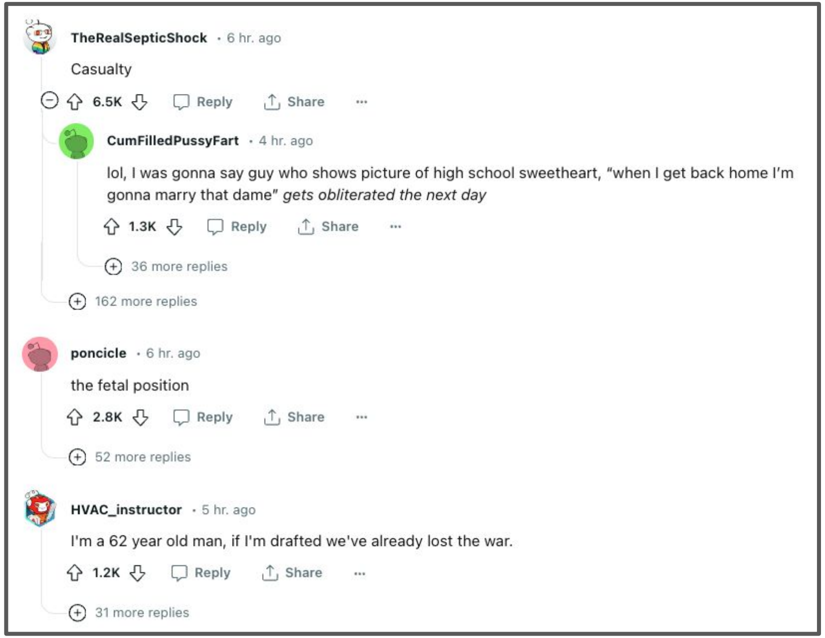
USC INFORMATION SCIENCES INSTITUTE

Carnegie Mellon University

SELECTSTAR

Reddit vs. Twitch conversations

Reddit



TheRealSepticShock · 6 hr. ago

Casualty

6.5K Reply Share ···

CumFilledPussyFart · 4 hr. ago

lol, I was gonna say guy who shows picture of high school sweetheart, "when I get back home I'm gonna marry that dame" *gets obliterated the next day*

1.3K Reply Share ···

36 more replies

162 more replies

poncicle · 6 hr. ago

the fetal position

2.8K Reply Share ···

52 more replies

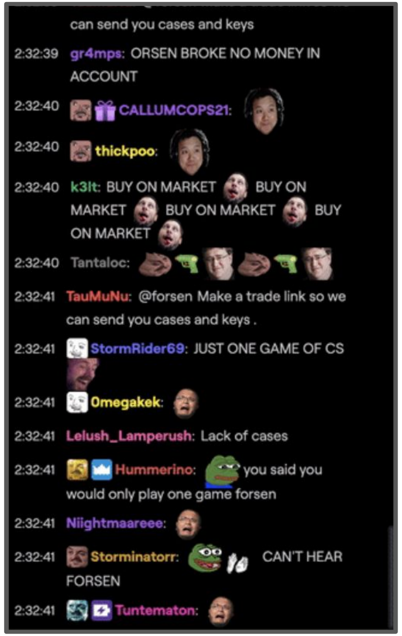
HVAC_instructor · 5 hr. ago

I'm a 62 year old man, if I'm drafted we've already lost the war.

1.2K Reply Share ···

31 more replies

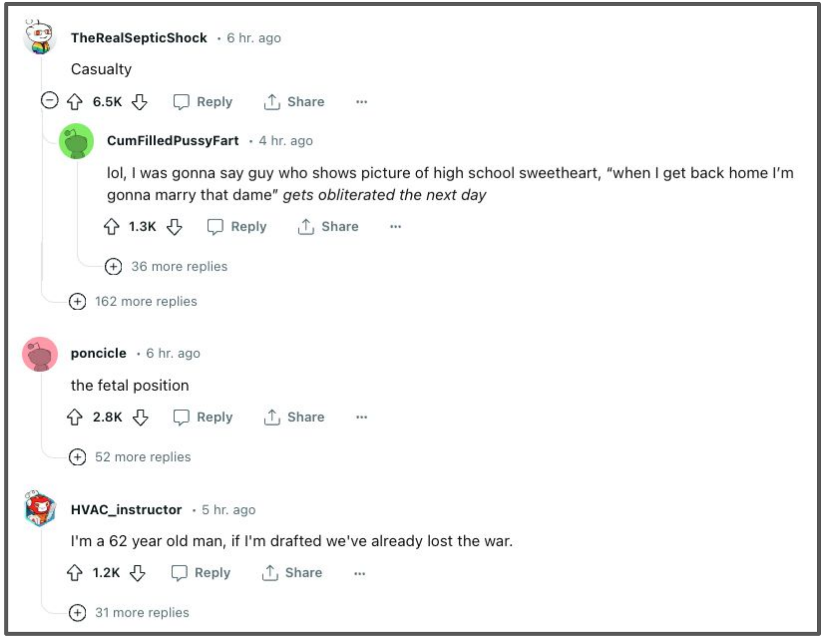
Twitch



can send you cases and keys
2:32:39 gr4mps: ORSEN BROKE NO MONEY IN ACCOUNT
2:32:40 CALLUMCOPS21:
2:32:40 thickpoo:
2:32:40 k3lt: BUY ON MARKET BUY ON MARKET BUY ON MARKET BUY ON MARKET
2:32:40 Tantaloc:
2:32:41 TauMuNu: @forsen Make a trade link so we can send you cases and keys .
2:32:41 StormRider69: JUST ONE GAME OF CS
2:32:41 Omegakek:
2:32:41 Lelush_Lamperush: Lack of cases
2:32:41 Hummerino: you said you would only play one game forsen
2:32:41 Niightmaareee:
2:32:41 Storminatorr: CAN'T HEAR FORSEN
2:32:41 Tuntematon:

Asynchronous vs. Synchronous conversations

Reddit (Asynchronous)



TheRealSepticShock · 6 hr. ago

Casualty

6.5K Reply Share ···

CumFilledPussyFart · 4 hr. ago

lol, I was gonna say guy who shows picture of high school sweetheart, “when I get back home I’m gonna marry that dame” *gets obliterated the next day*

1.3K Reply Share ···

36 more replies

162 more replies

poncicle · 6 hr. ago

the fetal position

2.8K Reply Share ···

52 more replies

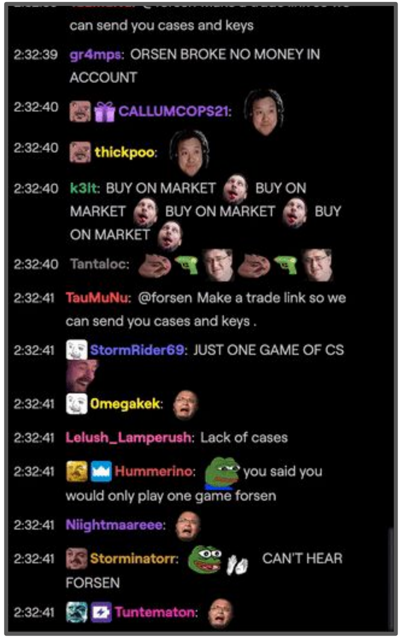
HVAC_instructor · 5 hr. ago

I'm a 62 year old man, if I'm drafted we've already lost the war.

1.2K Reply Share ···

31 more replies

Twitch (Synchronous)



can send you cases and keys

2:32:39 gr4mps: ORSEN BROKE NO MONEY IN ACCOUNT

2:32:40 CALLUMCOPS21:

2:32:40 thickpoo:

2:32:40 k3lt: BUY ON MARKET BUY ON MARKET BUY ON MARKET BUY ON MARKET

2:32:40 Tantaloc:

2:32:41 TauMuNu: @forsen Make a trade link so we can send you cases and keys .

2:32:41 StormRider69: JUST ONE GAME OF CS

2:32:41 Omegakek:

2:32:41 Lelush_Lamperush: Lack of cases

2:32:41 Hummerino: you said you would only play one game forsen

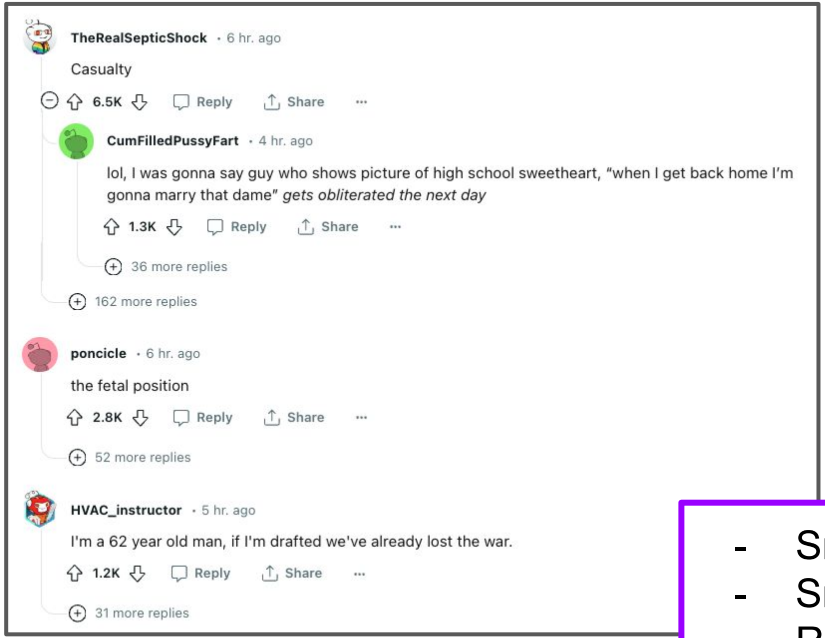
2:32:41 Niightmaareee:

2:32:41 Storminatorr: CAN'T HEAR FORSEN

2:32:41 Tuntematon:

Asynchronous vs. Synchronous conversations

Reddit (Asynchronous)



TheRealSepticShock · 6 hr. ago

Casualty

6.5K Reply Share ···

CumFilledPussyFart · 4 hr. ago

lol, I was gonna say guy who shows picture of high school sweetheart, "when I get back home I'm gonna marry that dame" *gets obliterated the next day*

1.3K Reply Share ···

36 more replies

162 more replies

poncicle · 6 hr. ago

the fetal position

2.8K Reply Share ···

52 more replies

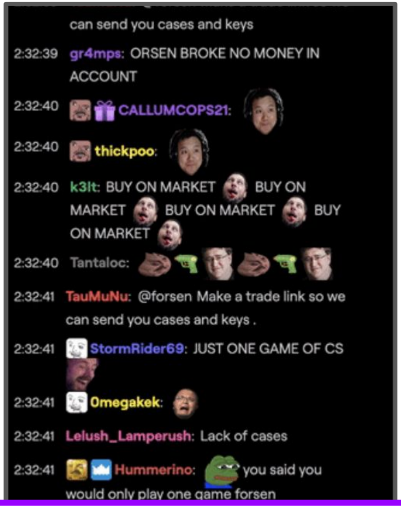
HVAC_instructor · 5 hr. ago

I'm a 62 year old man, if I'm drafted we've already lost the war.

1.2K Reply Share ···

31 more replies

Twitch (Synchronous)



can send you cases and keys

2:32:39 gr4mps: ORSEN BROKE NO MONEY IN ACCOUNT

2:32:40 CALLUMCOPS21:

2:32:40 thickpoo:

2:32:40 k3lt: BUY ON MARKET BUY ON MARKET BUY ON MARKET BUY ON MARKET

2:32:40 Tantaloc:

2:32:41 TauMuNu: @forsen Make a trade link so we can send you cases and keys .

2:32:41 StormRider69: JUST ONE GAME OF CS

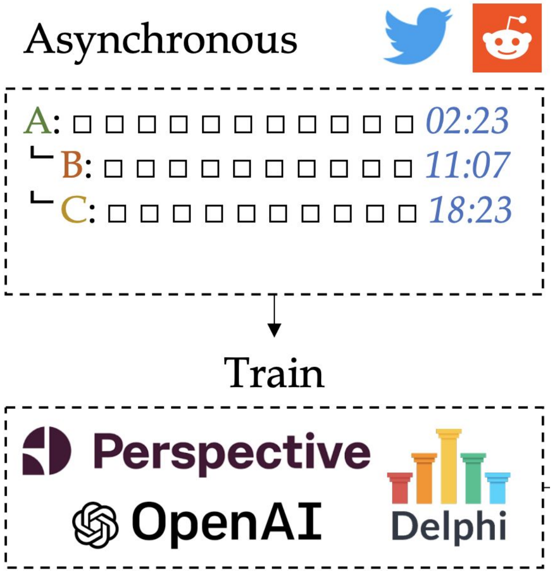
2:32:41 Omegakek:

2:32:41 Lelush_Lamperush: Lack of cases

2:32:41 Hummerino: you said you would only play one game forsen

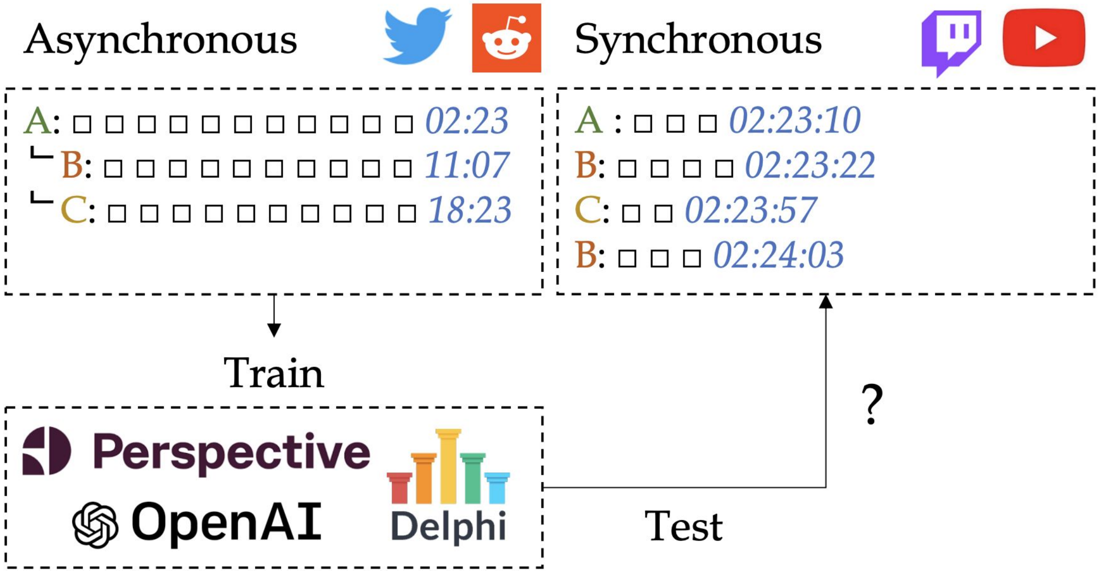
- Smaller message length.
- Smaller temporal gap between chats.
- Relationships between chats are less clearly defined.

Existing Toxicity Detection Models


Asynchronous
A: □ □ □ □ □ □ □ □ □ □ □ 02:23
B: □ □ □ □ □ □ □ □ □ □ 11:07
C: □ □ □ □ □ □ □ □ □ □ 18:23
Train
Perspective
OpenAI
Delphi

Existing models primarily trained on toxicity-labeled data in the **asynchronous** domain.

What if we use such models on **Synchronous** domain?



Asynchronous

A: □ □ □ □ □ □ □ □ □ □ □ *02:23*
└ B: □ □ □ □ □ □ □ □ □ □ *11:07*
└ C: □ □ □ □ □ □ □ □ □ □ *18:23*

Synchronous

A : □ □ □ *02:23:10*
B: □ □ □ □ *02:23:22*
C: □ □ *02:23:57*
B: □ □ □ *02:24:03*

Train

Test

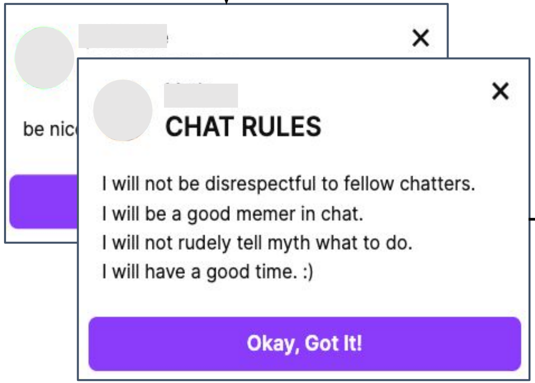
?

Can these models perform well in
detecting violations also in the **synchronous** domain?

Data Construction

Twitch Top 200 Streamers

Gather chat rules



[redacted]

be nic

[redacted]

CHAT RULES

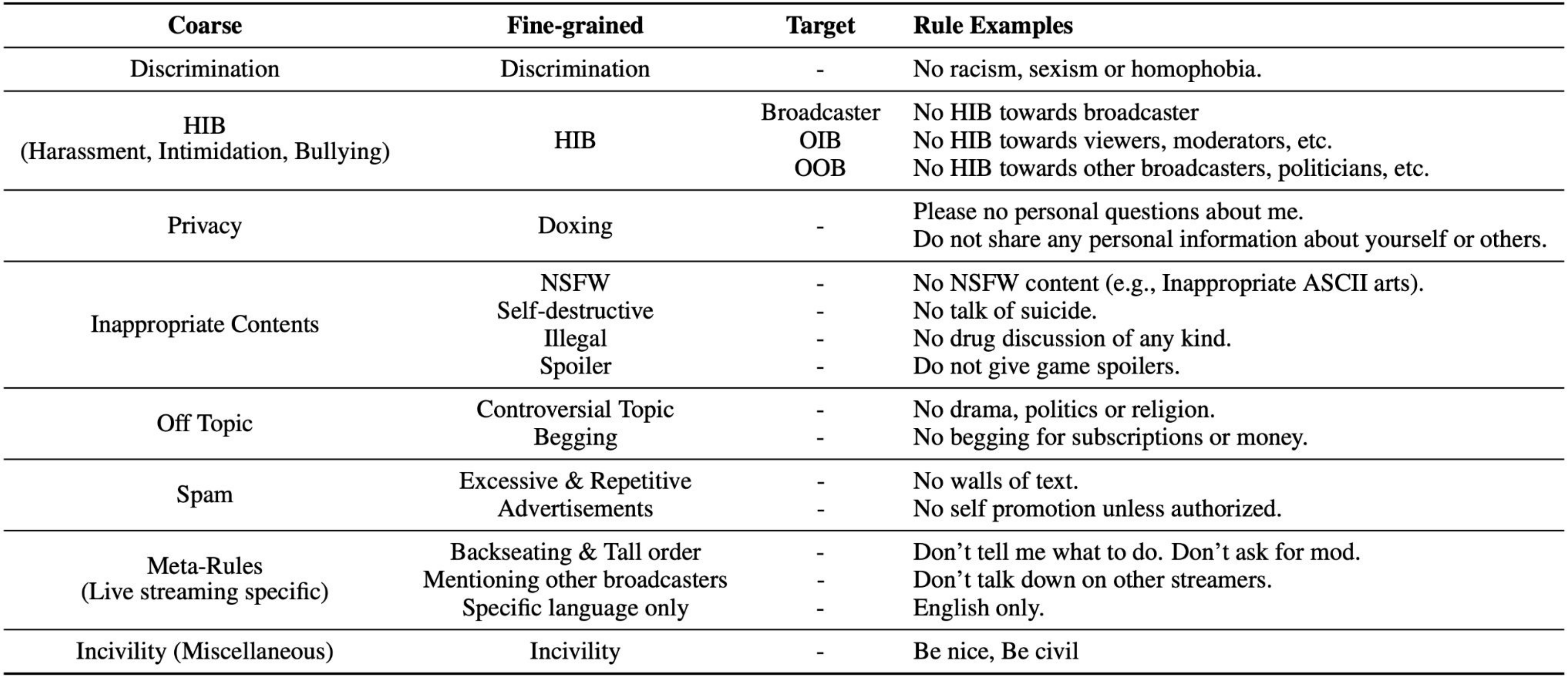
I will not be disrespectful to fellow chatters.
I will be a good memer in chat.
I will not rudely tell myth what to do.
I will have a good time. :)

Define Norms

Map rules

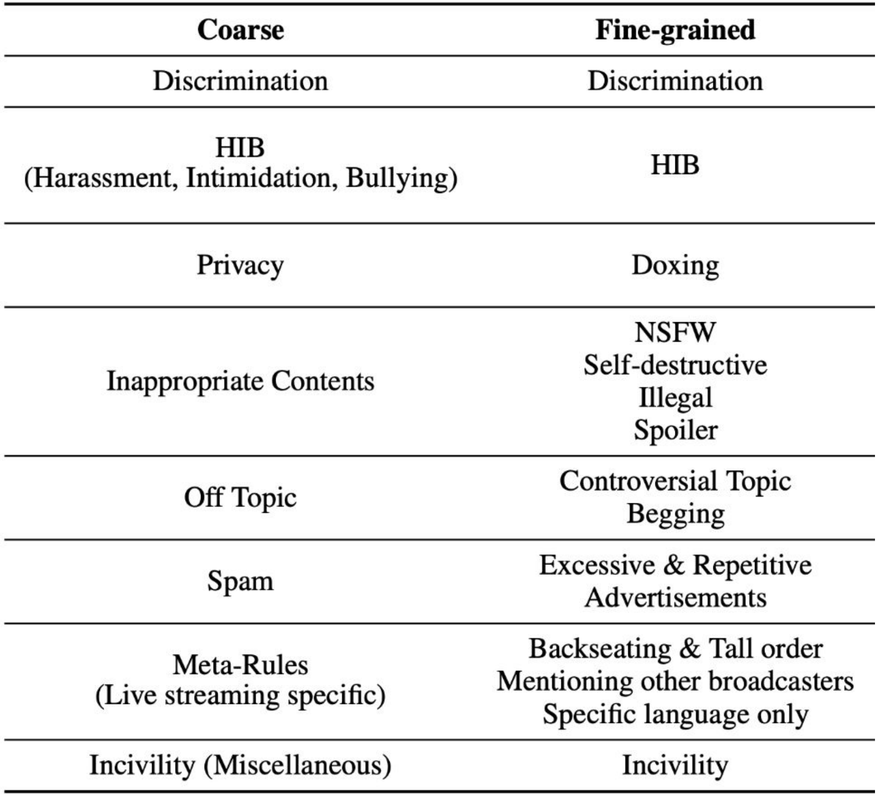
Norm	Rule Examples
Incivility	Be Nice
Discrimination	No Racism
…	…

Violation norms in **Twitch**



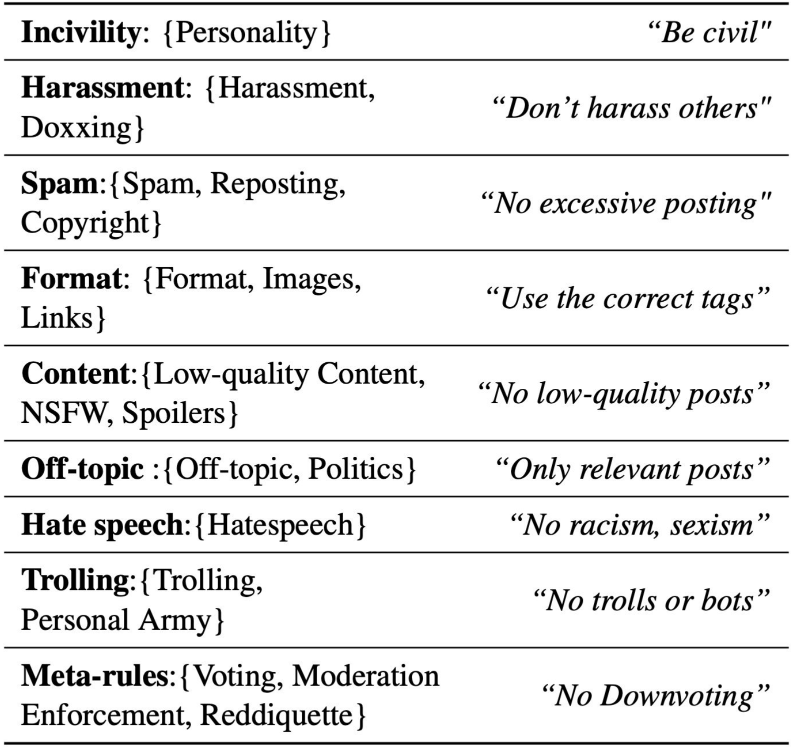
Coarse	Fine-grained	Target	Rule Examples
Discrimination	Discrimination	-	No racism, sexism or homophobia.
HIB (Harassment, Intimidation, Bullying)	HIB	Broadcaster OIB OOB	No HIB towards broadcaster No HIB towards viewers, moderators, etc. No HIB towards other broadcasters, politicians, etc.
Privacy	Doxing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

Violation norms in Twitch vs Reddit



Coarse	Fine-grained
Discrimination	Discrimination
HIB (Harassment, Intimidation, Bullying)	HIB
Privacy	Doxing
Inappropriate Contents	NSFW Self-destructive Illegal Spoiler
Off Topic	Controversial Topic Begging
Spam	Excessive & Repetitive Advertisements
Meta-Rules (Live streaming specific)	Backseating & Tall order Mentioning other broadcasters Specific language only
Incivility (Miscellaneous)	Incivility

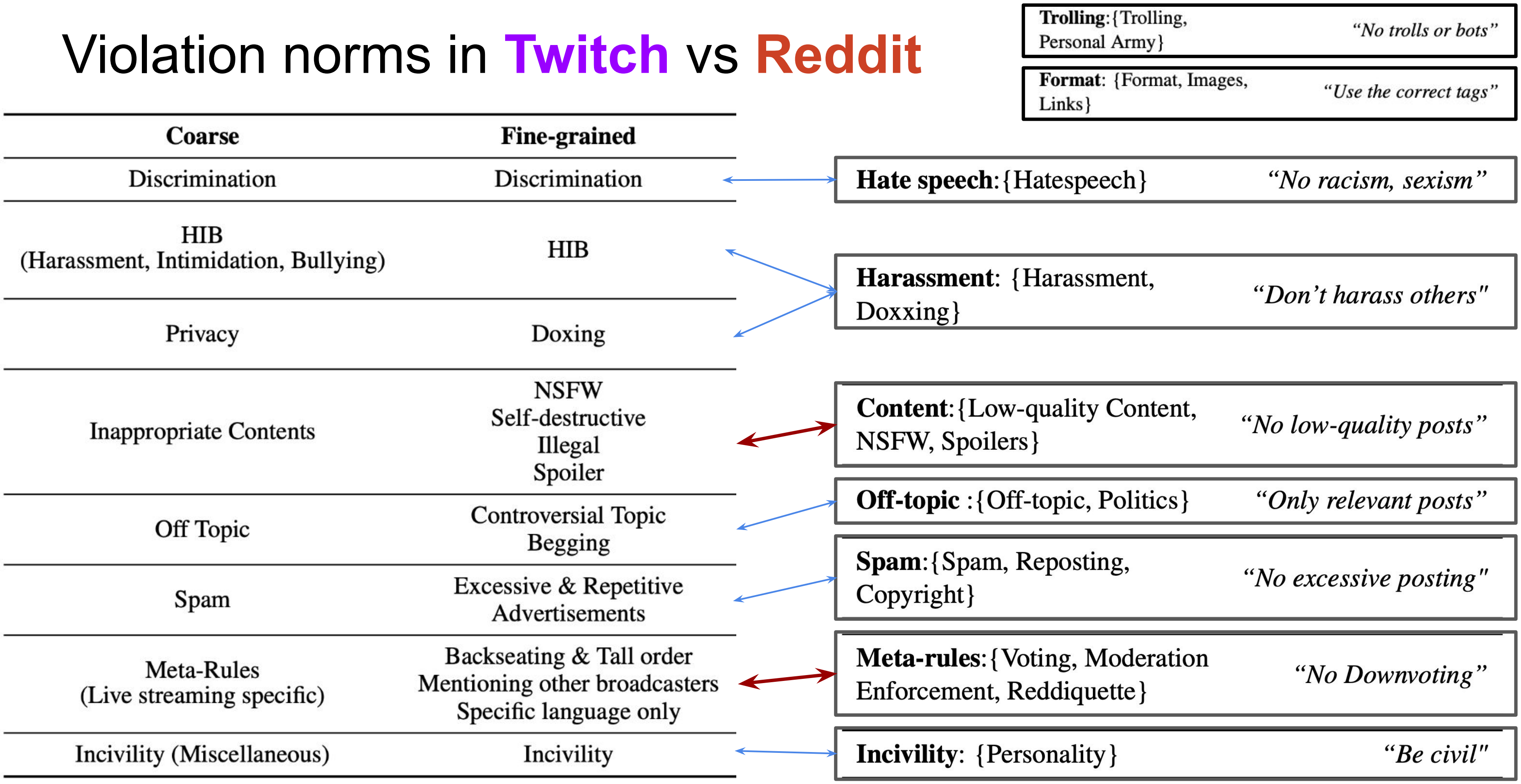
Reddit (NormVio, Park et al., 2021)



Incivility: {Personality}	*"Be civil"*
Harassment: {Harassment, Doxxing}	*"Don't harass others"*
Spam:{Spam, Reposting, Copyright}	*"No excessive posting"*
Format: {Format, Images, Links}	*"Use the correct tags"*
Content:{Low-quality Content, NSFW, Spoilers}	*"No low-quality posts"*
Off-topic :{Off-topic, Politics}	*"Only relevant posts"*
Hate speech:{Hatespeech}	*"No racism, sexism"*
Trolling:{Trolling, Personal Army}	*"No trolls or bots"*
Meta-rules:{Voting, Moderation Enforcement, Reddiquette}	*"No Downvoting"*

Detecting Community Sensitive Norm Violations in Online Conversations., **Park** et al., EMNLP 2021 Findings

Violation norms in Twitch vs Reddit



Coarse	Fine-grained
Discrimination	Discrimination
HIB (Harassment, Intimidation, Bullying)	HIB
Privacy	Doxing
Inappropriate Contents	NSFW Self-destructive Illegal Spoiler
Off Topic	Controversial Topic Begging
Spam	Excessive & Repetitive Advertisements
Meta-Rules (Live streaming specific)	Backseating & Tall order Mentioning other broadcasters Specific language only
Incivility (Miscellaneous)	Incivility

Trolling:{Trolling, Personal Army} *"No trolls or bots"*

Format: {Format, Images, Links} *"Use the correct tags"*

Hate speech:{Hatespeech} *"No racism, sexism"*

Harassment: {Harassment, Doxxing} *"Don't harass others"*

Content:{Low-quality Content, NSFW, Spoilers} *"No low-quality posts"*

Off-topic :{Off-topic, Politics} *"Only relevant posts"*

Spam:{Spam, Reposting, Copyright} *"No excessive posting"*

Meta-rules:{Voting, Moderation Enforcement, Reddiquette} *"No Downvoting"*

Incivility: {Personality} *"Be civil"*

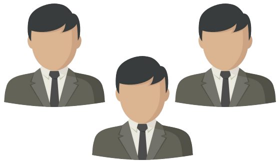
Detecting Community Sensitive Norm Violations in Online Conversations., **Park** et al., EMNLP 2021 Findings

Data Construction

Twitch Top 200 Streamers

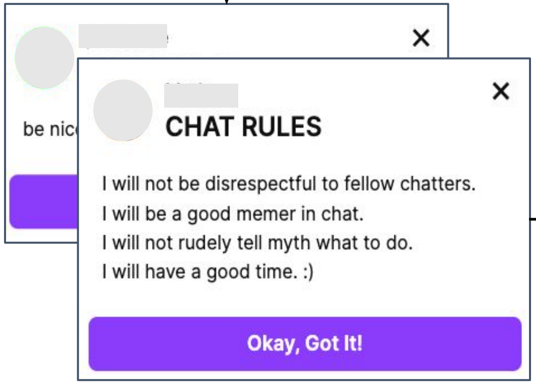
Twitch API & IRC

Monitor Moderation Events (Ban, Timeout, Delete)



Annotators

Gather chat rules



[redacted]

be nic

[redacted]

CHAT RULES

I will not be disrespectful to fellow chatters.
I will be a good memer in chat.
I will not rudely tell myth what to do.
I will have a good time. :)

Define Norms

Map rules

Norm	Rule Examples
Incivility	Be Nice
Discrimination	No Racism
…	…

Data Construction

Twitch Top 200 Streamers

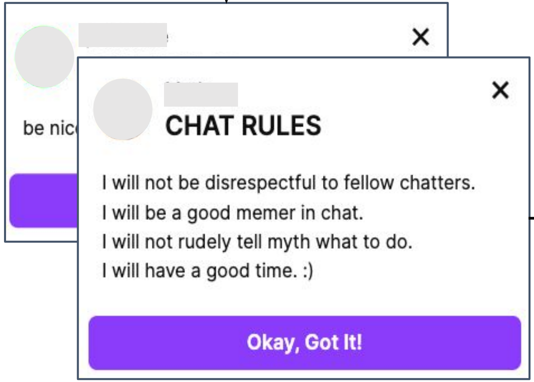
Twitch API & IRC

Monitor Moderation Events (Ban, Timeout, Delete)

Annotators

Gather chat rules

Stage 1	Moderated Message
Stage 2	Moderated Message + Chat Context
Stage 3	Moderated Message + Chat / Video Context



[redacted]

be nic

[redacted]

CHAT RULES

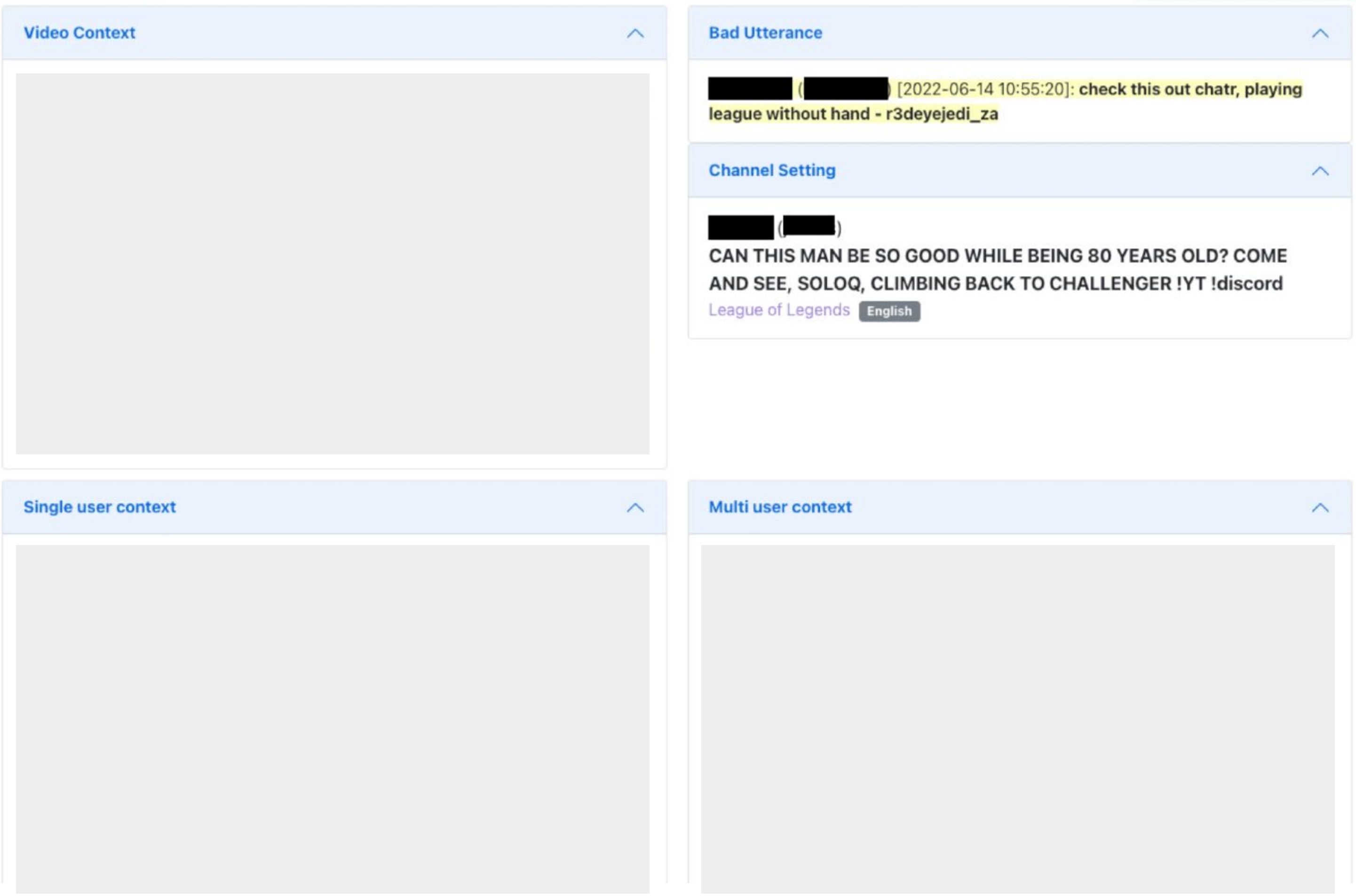
I will not be disrespectful to fellow chatters.
I will be a good memer in chat.
I will not rudely tell myth what to do.
I will have a good time. :)

Define Norms

Map rules

Norm	**Rule Examples**
Incivility	Be Nice
Discrimination	No Racism
…	…

Annotation Interface – Stage 1: Moderated Message



Video Context

Bad Utterance

[redacted] ([redacted]) [2022-06-14 10:55:20]: **check this out chatr, playing league without hand - r3deyejedi_za**

Channel Setting

[redacted] ([redacted])

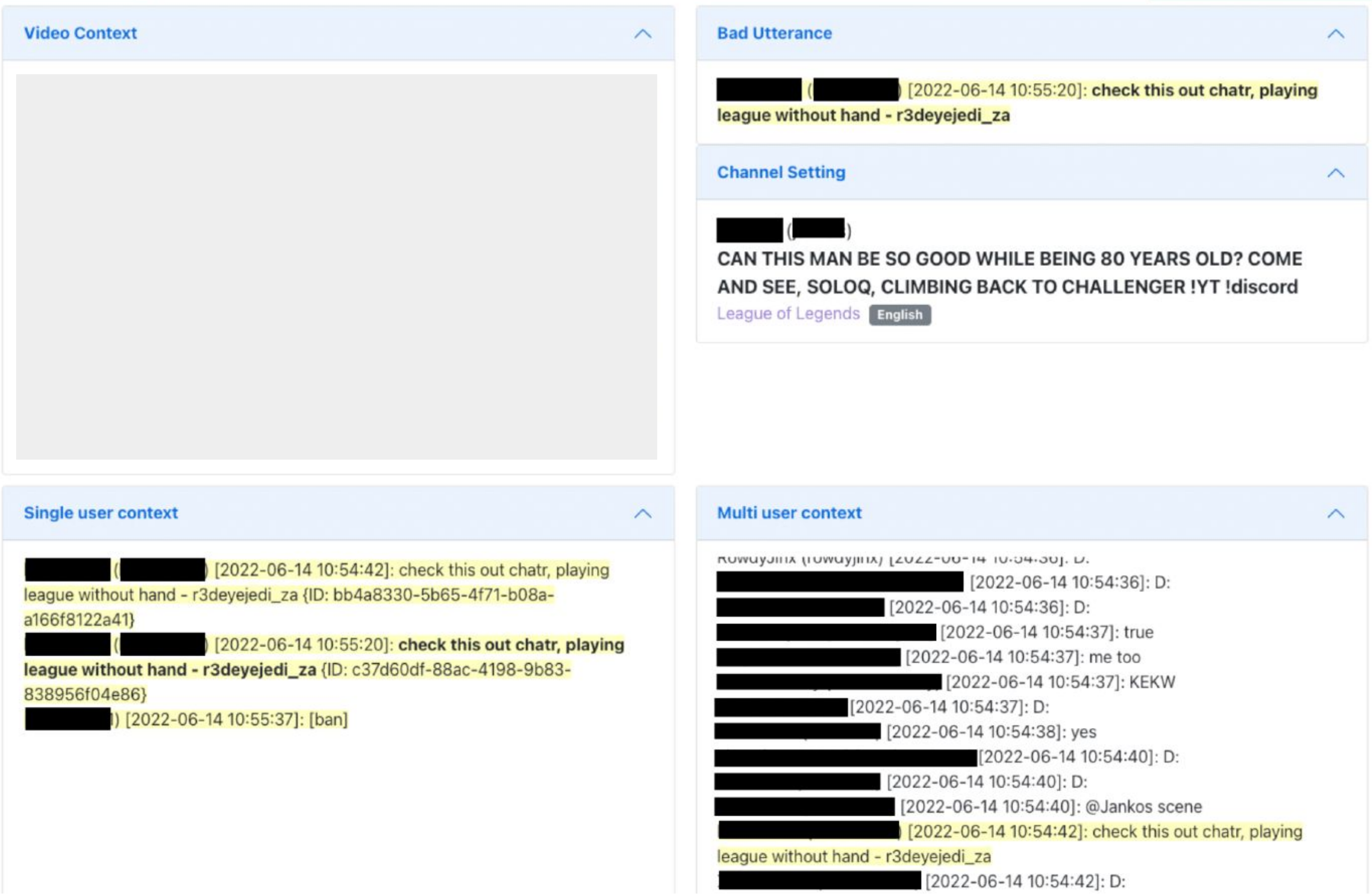
CAN THIS MAN BE SO GOOD WHILE BEING 80 YEARS OLD? COME AND SEE, SOLOQ, CLIMBING BACK TO CHALLENGER !YT !discord

League of Legends English

Single user context

Multi user context

Annotation Interface – Stage 2: Moderated Message + Chat Context



Video Context

Bad Utterance

[redacted] ([redacted]) [2022-06-14 10:55:20]: **check this out chatr, playing league without hand - r3deyejedi_za**

Channel Setting

[redacted] ([redacted])
CAN THIS MAN BE SO GOOD WHILE BEING 80 YEARS OLD? COME AND SEE, SOLOQ, CLIMBING BACK TO CHALLENGER !YT !discord
League of Legends English

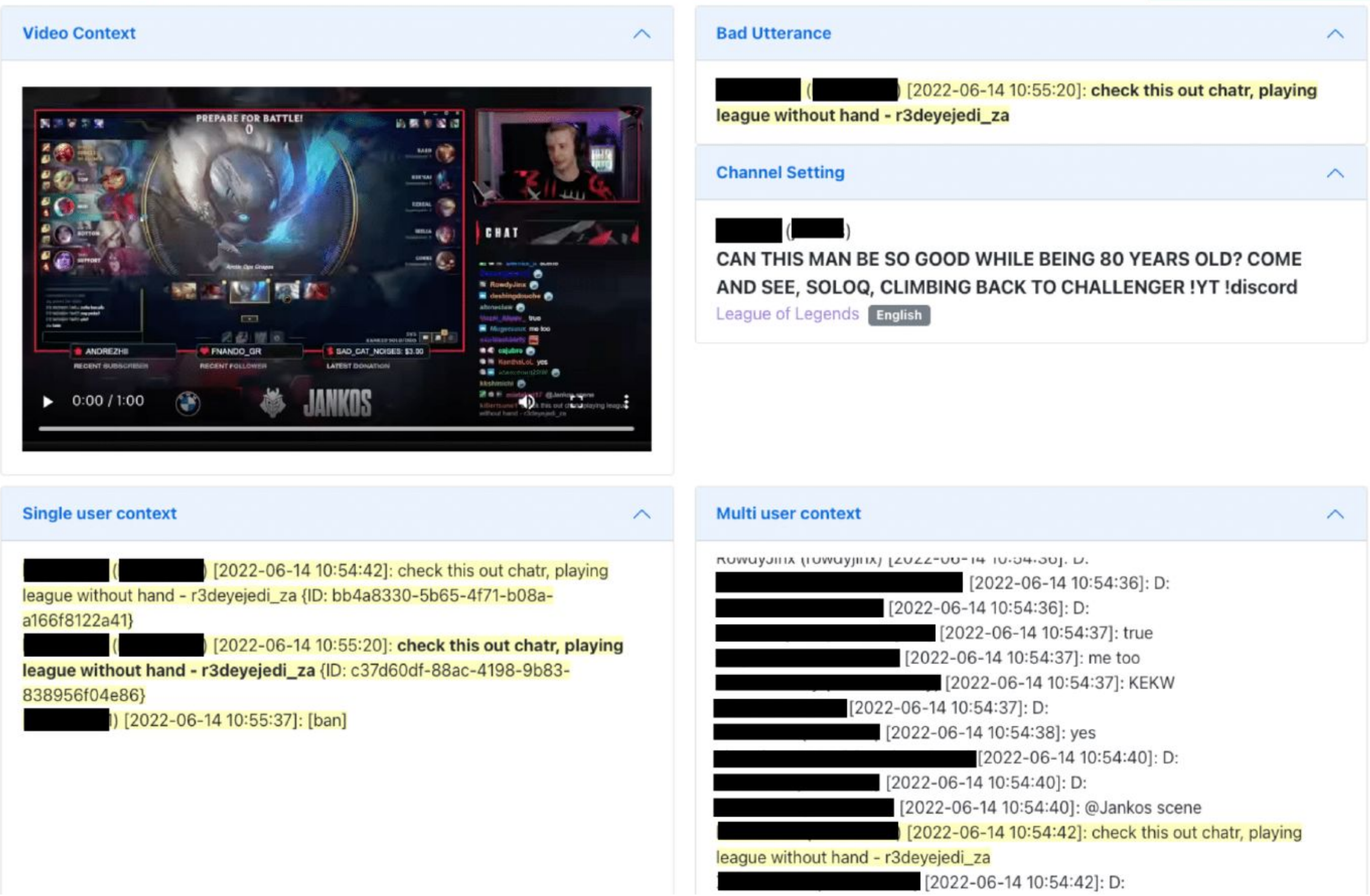
Single user context

[redacted] ([redacted]) [2022-06-14 10:54:42]: check this out chatr, playing league without hand - r3deyejedi_za {ID: bb4a8330-5b65-4f71-b08a-a166f8122a41}
[redacted] ([redacted]) [2022-06-14 10:55:20]: **check this out chatr, playing league without hand - r3deyejedi_za** {ID: c37d60df-88ac-4198-9b83-838956f04e86}
[redacted]) [2022-06-14 10:55:37]: [ban]

Multi user context

RowdyJinx (rowdyjinx) [2022-06-14 10:54:36]: D:
[redacted] [2022-06-14 10:54:36]: D:
[redacted] [2022-06-14 10:54:36]: D:
[redacted] [2022-06-14 10:54:37]: true
[redacted] [2022-06-14 10:54:37]: me too
[redacted] [2022-06-14 10:54:37]: KEKW
[redacted] [2022-06-14 10:54:37]: D:
[redacted] [2022-06-14 10:54:38]: yes
[redacted] [2022-06-14 10:54:40]: D:
[redacted] [2022-06-14 10:54:40]: D:
[redacted] [2022-06-14 10:54:40]: @Jankos scene
[redacted] [2022-06-14 10:54:42]: check this out chatr, playing league without hand - r3deyejedi_za
[redacted] [2022-06-14 10:54:42]: D:

Annotation Interface – Stage 3: Moderated Message + Chat/Video Context



Video Context

Bad Utterance

[redacted] ([redacted]) [2022-06-14 10:55:20]: **check this out chatr, playing league without hand - r3deyejedi_za**

Channel Setting

[redacted] ([redacted])

CAN THIS MAN BE SO GOOD WHILE BEING 80 YEARS OLD? COME AND SEE, SOLOQ, CLIMBING BACK TO CHALLENGER !YT !discord

League of Legends English

Single user context

[redacted] ([redacted]) [2022-06-14 10:54:42]: check this out chatr, playing league without hand - r3deyejedi_za {ID: bb4a8330-5b65-4f71-b08a-a166f8122a41}

[redacted] ([redacted]) [2022-06-14 10:55:20]: **check this out chatr, playing league without hand - r3deyejedi_za** {ID: c37d60df-88ac-4198-9b83-838956f04e86}

[redacted]) [2022-06-14 10:55:37]: [ban]

Multi user context

RowdyJinx (rowdyjinx) [2022-06-14 10:54:36]: D:

[redacted] [2022-06-14 10:54:36]: D:

[redacted] [2022-06-14 10:54:36]: D:

[redacted] [2022-06-14 10:54:37]: true

[redacted] [2022-06-14 10:54:37]: me too

[redacted] [2022-06-14 10:54:37]: KEKW

[redacted] [2022-06-14 10:54:37]: D:

[redacted] [2022-06-14 10:54:38]: yes

[redacted] [2022-06-14 10:54:40]: D:

[redacted] [2022-06-14 10:54:40]: D:

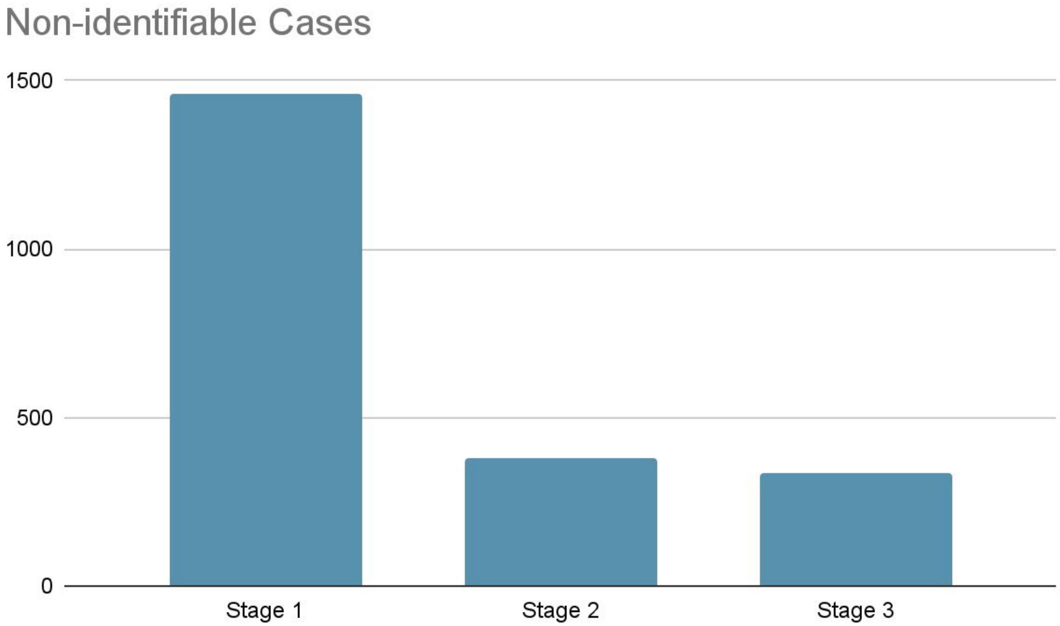
[redacted] [2022-06-14 10:54:40]: @Jankos scene

[redacted] [2022-06-14 10:54:42]: check this out chatr, playing league without hand - r3deyejedi_za

[redacted] [2022-06-14 10:54:42]: D:

Data Statistics

Number of non-identifiable cases

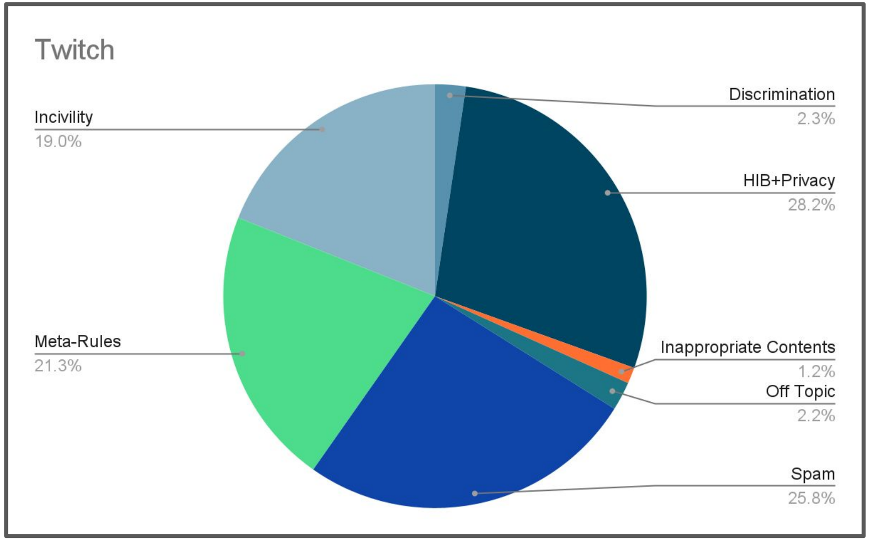


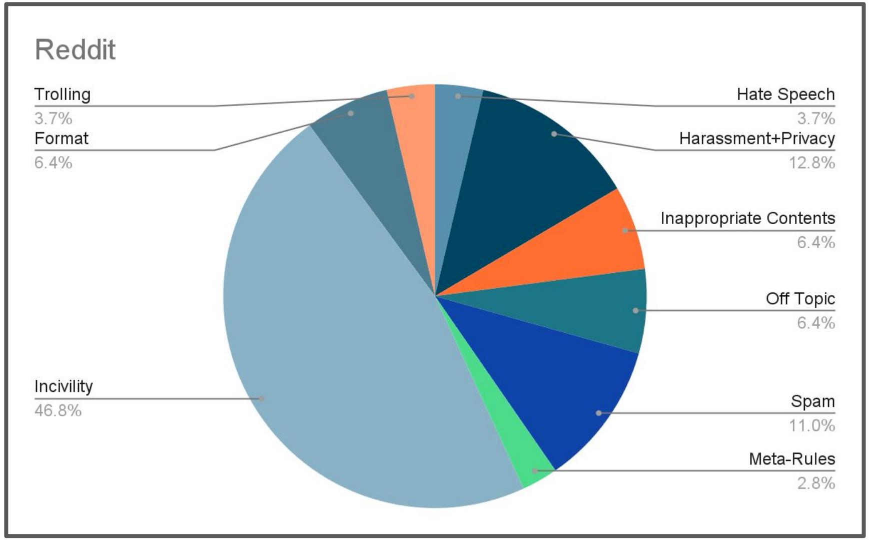
Providing context will lessen the ambiguity of the violation.

Data Statistics

Twitch vs. Reddit

Reddit (NormVio, Park et al., 2021)





Similarity

(1) **Harassment** (HIB) and **Incivility** take up a large portion.

(2) **Off-topic, Inappropriate contents, Privacy** exist but less enforced.

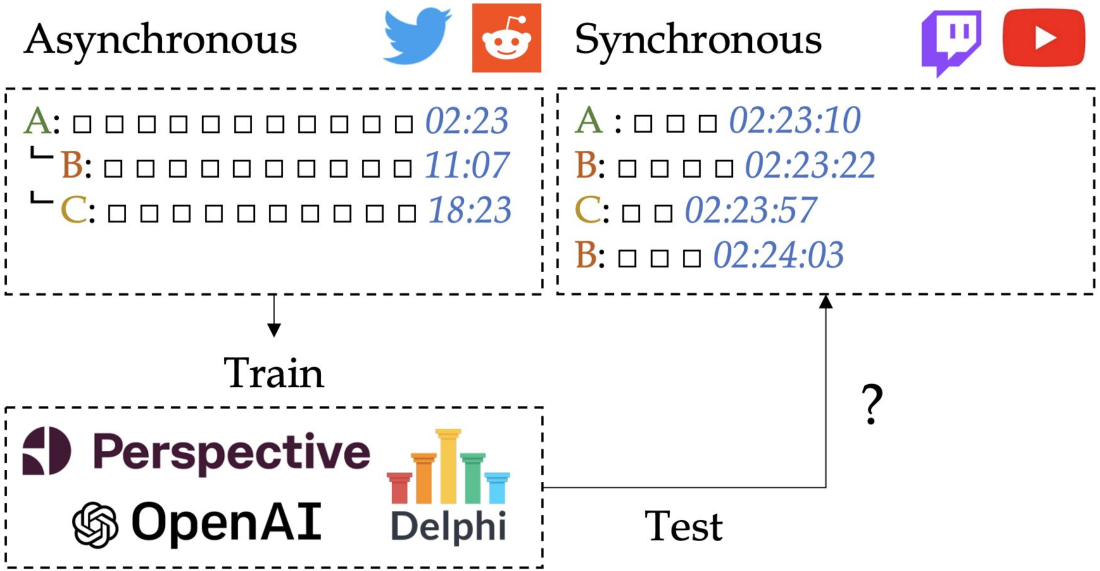
Difference

(1) **Spam**, **Meta-Rules** cover significantly higher portion in Twitch.

(2) Twitch streamers are less concerned about **contents**.

Performance of existing model endpoints.

Recap



Asynchronous

A: □ □ □ □ □ □ □ □ □ □ □ 02:23

└ B: □ □ □ □ □ □ □ □ □ □ 11:07

└ C: □ □ □ □ □ □ □ □ □ □ 18:23

Synchronous

A : □ □ □ 02:23:10

B: □ □ □ □ 02:23:22

C: □ □ 02:23:57

B: □ □ □ 02:24:03

Train

Perspective

OpenAI

Delphi

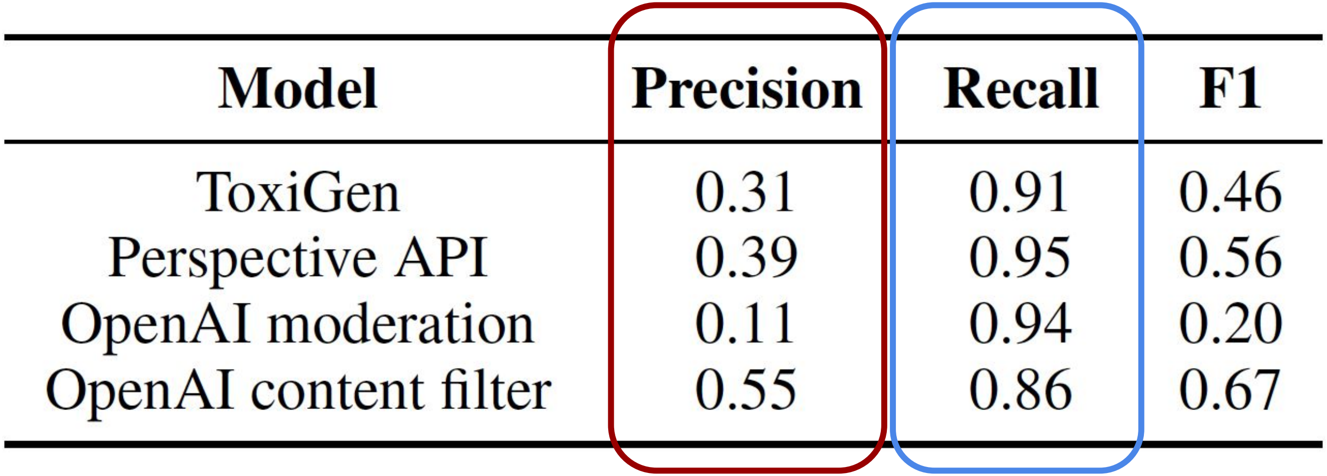
Test

?

Can these models perform well in
detecting violations also in the **synchronous** domain?

Performance of existing model endpoints.

Binary classification on (Discrimination & HIB)

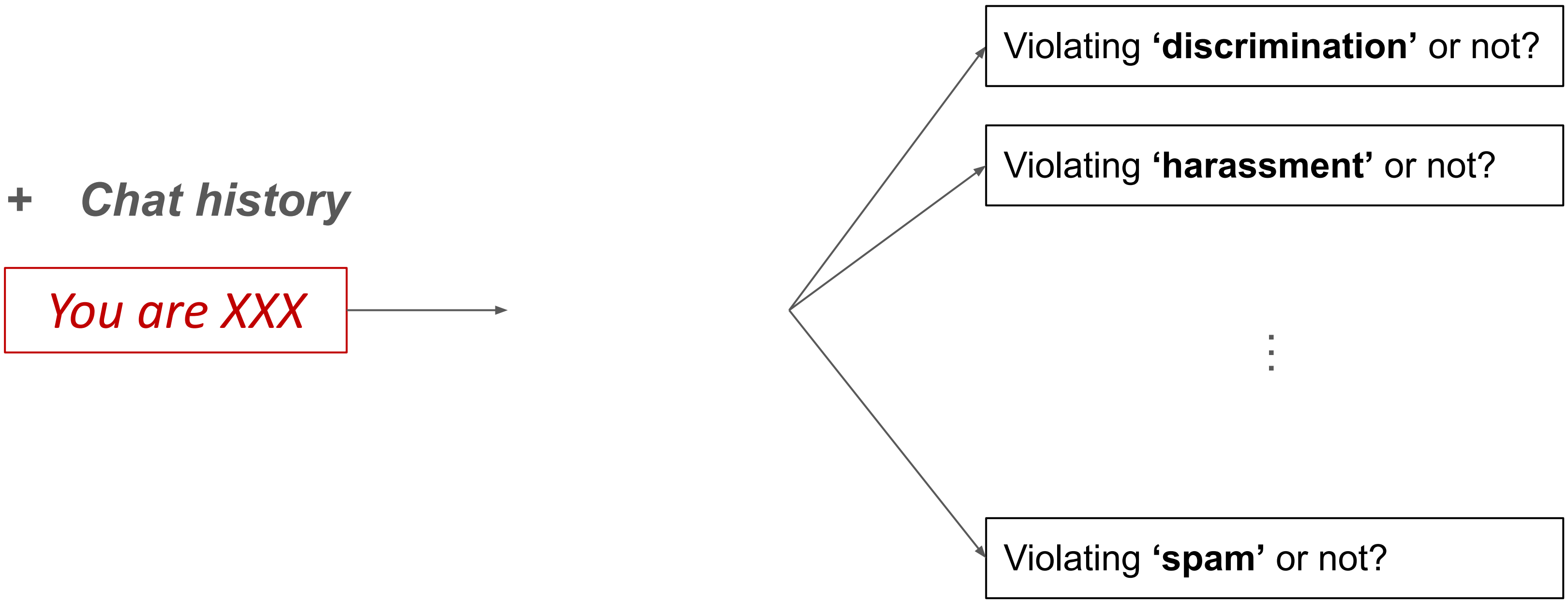


Model	Precision	Recall	F1
ToxiGen	0.31	0.91	0.46
Perspective API	0.39	0.95	0.56
OpenAI moderation	0.11	0.94	0.20
OpenAI content filter	0.55	0.86	0.67

Existing models do not frequently produce false positives (**high recall**).
However, they perform poorly in detecting toxic messages (**low precision**).

Norm Classification

Binary classification for each category



+ *Chat history*

You are XXX

Violating **‘discrimination’** or not?

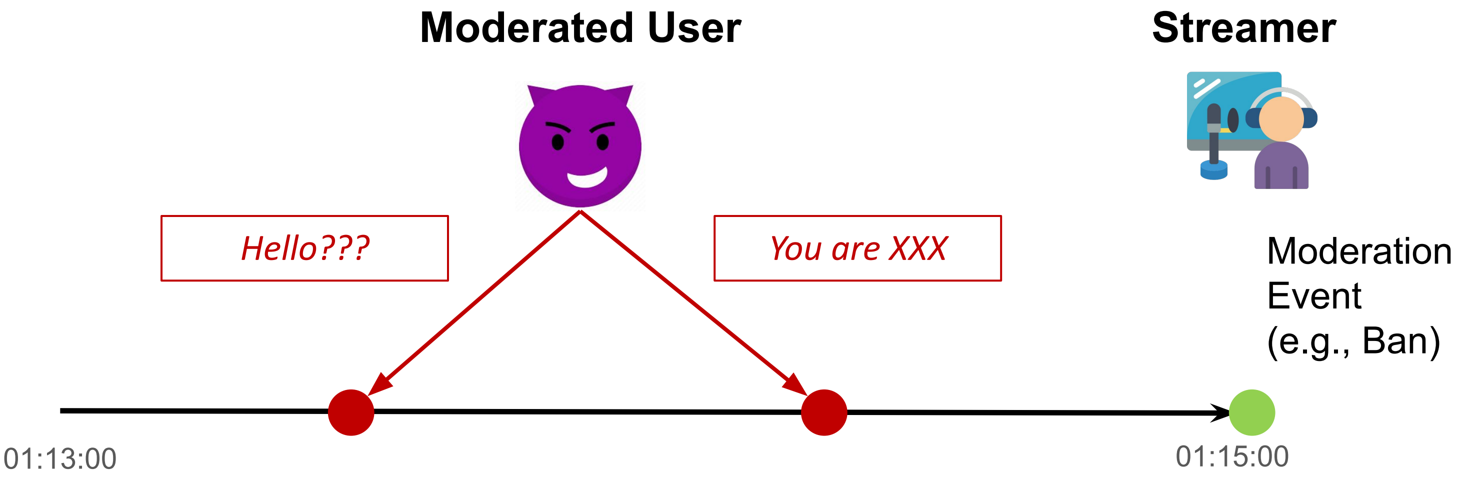
Violating **‘harassment’** or not?

⋮

Violating **‘spam’** or not?

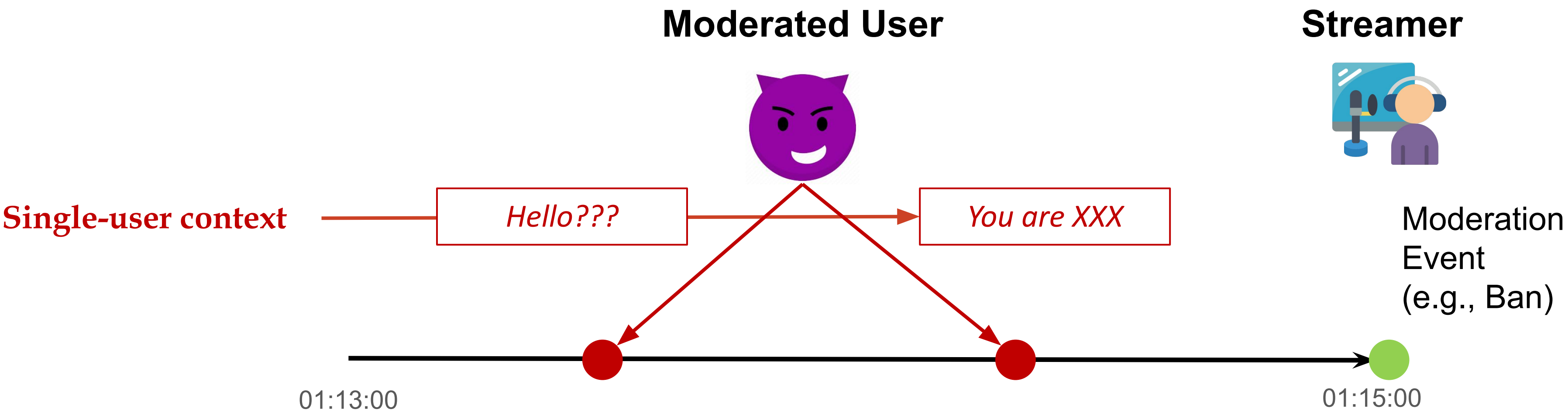
Norm Classification

Context



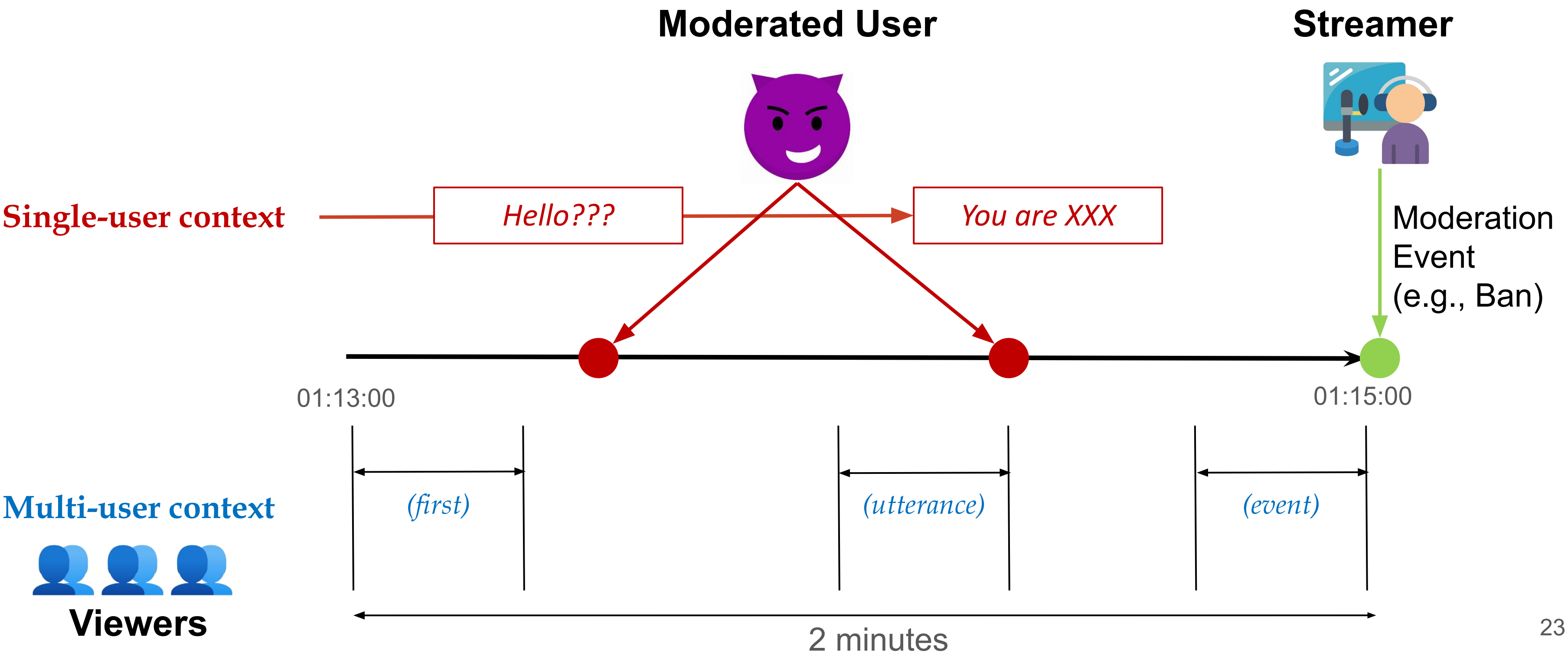
Norm Classification

Single-user Context



Norm Classification

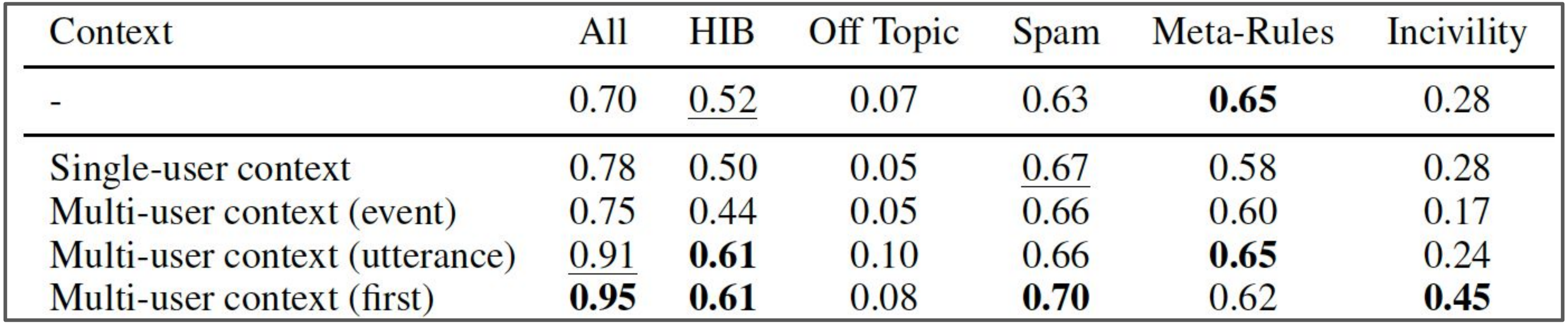
Multi-user Context



Norm Classification

Performance

All: Any norm violated? or not?



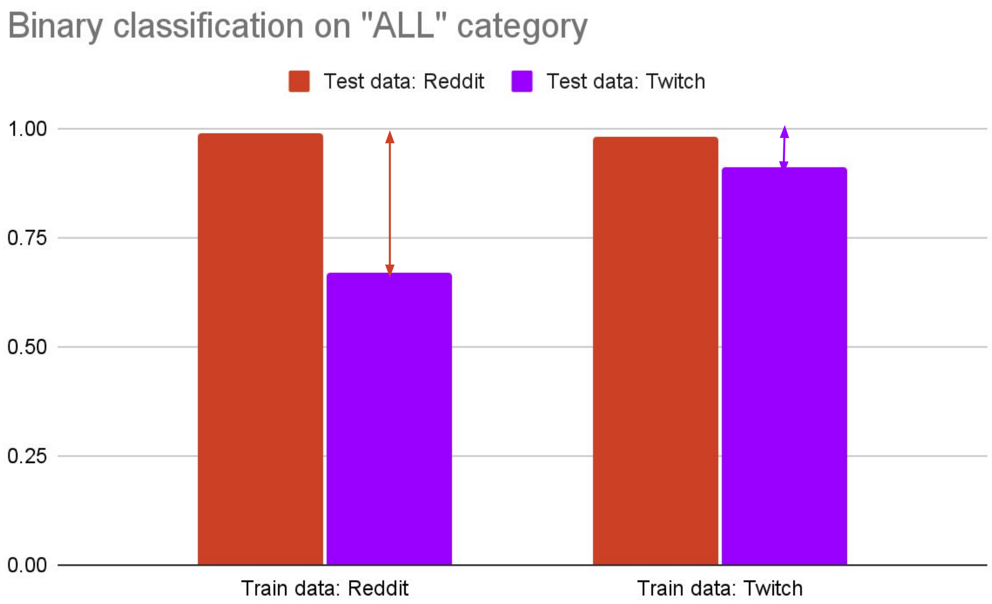
Context	All	HIB	Off Topic	Spam	Meta-Rules	Incivility
-	0.70	0.52	0.07	0.63	**0.65**	0.28
Single-user context	0.78	0.50	0.05	0.67	0.58	0.28
Multi-user context (event)	0.75	0.44	0.05	0.66	0.60	0.17
Multi-user context (utterance)	0.91	**0.61**	0.10	0.66	**0.65**	0.24
Multi-user context (first)	**0.95**	**0.61**	0.08	**0.70**	0.62	**0.45**

Context matters a lot for training better norm classification model!

- Chats from other users (multi-user context) help determine the toxicity
- However, temporal gap between event and the actual offending chat may be substantial

Norm Classification

Train on Reddit → Test on Twitch / Train on Twitch → Test on Reddit



Model trained on Twitch **shows better generalizability** than model trained on Reddit

Conclusion

1. Norm violations vary between synchronous platforms like Twitch and asynchronous ones like Reddit

2. **Context is crucial** for improving models that detect chat norm violations.

3. A model trained on Twitch data may be more effective in identifying real-world norm violations.

Q&A

Dong-Ho Lee

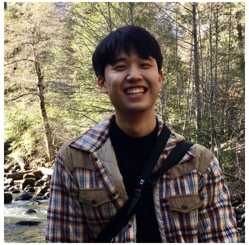
Jihyung Moon

Sungjoon Park

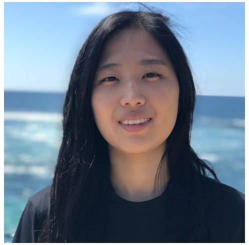
softly.ai

USC INFORMATION SCIENCES INSTITUTE

Carnegie Mellon University


Hyundong Cho

Woojeong Jin


Chanyoung Park

Jonathan May

Jay Pujara

Data labeling sponsor: SELECTSTAR