UNIVERSITÄT LIECHTENSTEIN

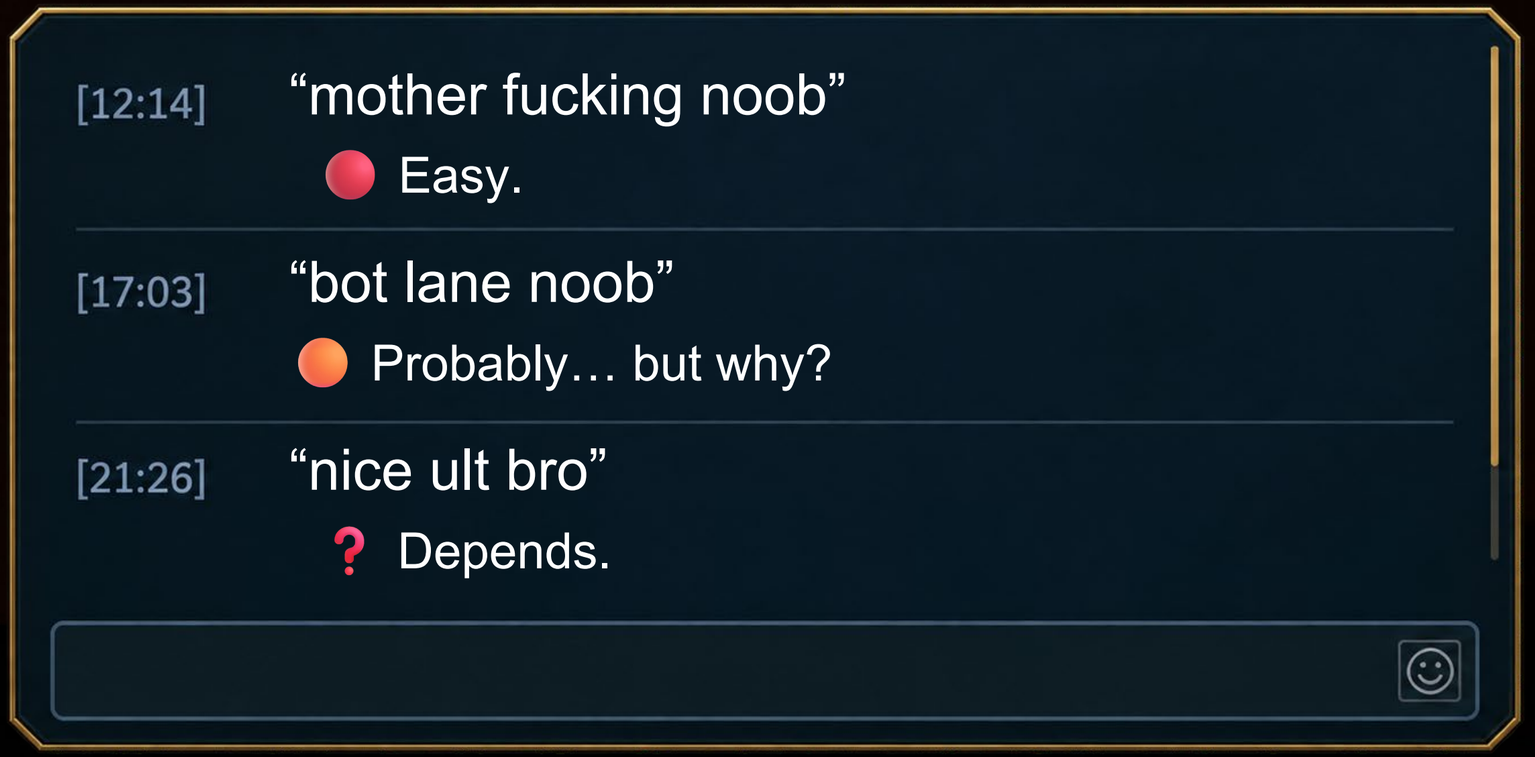
“BOT LANE NOOB” TOWARDS DEPLOYMENT OF NLP-BASED TOXICITY DETECTORS IN VIDEO GAMES

ESORICS 2026

Jonas Ave, **Irdin Pekaric** ,
Matthias Frohner and Giovanni Apruzzese


WARNING
OFFENSIVE LANGUAGE

IS THIS TOXIC?



Gaming toxicity is contextual.

Slang. Sarcasm. Game - specific language.

WHY NOT USE AN EXISTING TOXIC



General Language ≠ Gaming Language

A detector trained outside gaming may understand the words, but not the game.

SYSTEMATIC LITERATURE REVIEW

Have ML/NLP methods actually been developed to detect toxicity during gameplay?

SEARCH
1,039
papers

→

SCREENING
50
candidates

→

FULL- TEXT REVIEW
15
ML/ NLP works

→

RELEVANT WORK
<2%
of papers

Only 15 papers

And even these had practical limitations.

Not during gameplay

Twitch. Youtube.
Forums. Reviews.

Data unavailable

No public dataset /
limited reproducibility

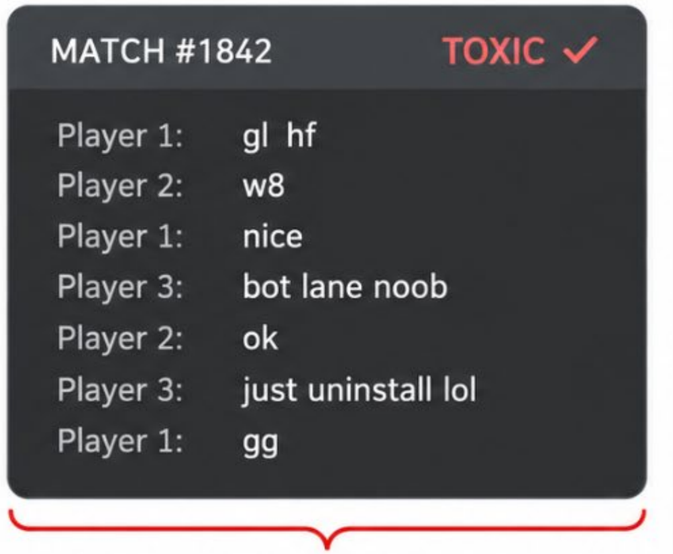
Limited scope

Specific toxicity type,
language, or game

Lots of research on toxicity. Very little built for the actual match.

DATA WAS THE BOTTLENECK

Existing Tribunal Data



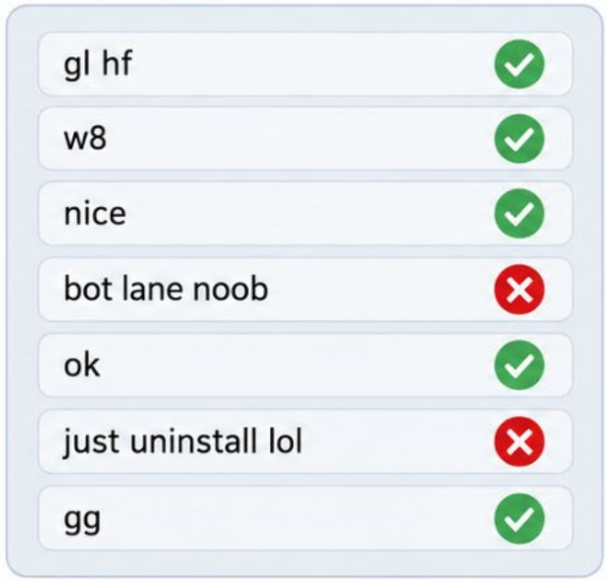
TOXIC MATCH

One label for the whole match

But which message is toxic?

Match - level labels are not enough

What We Actually Need



MESSAGE - LEVEL GROUND TRUTH

One label for each message

You cannot train a message detector with only match - level labels.

LABELLING & DATASET

8 Expert LoL Players



6-20 years of gaming experience

Initial Annotation

bot lane noob	toxic
nice ult bro	non-toxic
just uninstall lol	toxic

8 annotators independently labelled the same first 5,000 messages

Consensus Rule



annotators say toxic

↓

TOXIC

A message is considered toxic if at least two annotators label it toxic

Review & Additional Labelling



3 annotators reviewed decisions, discussed borderline cases, and labelled 10,999 additional messages

1,398
toxic
8.74%

13,773
non - toxic
86.09%

828
non - English
5.17%

Toxic Examples

toxic

bot lane noob

just uninstall lol

Non - Toxic Examples

glhf

ok

gg

Filtered Out

non - English

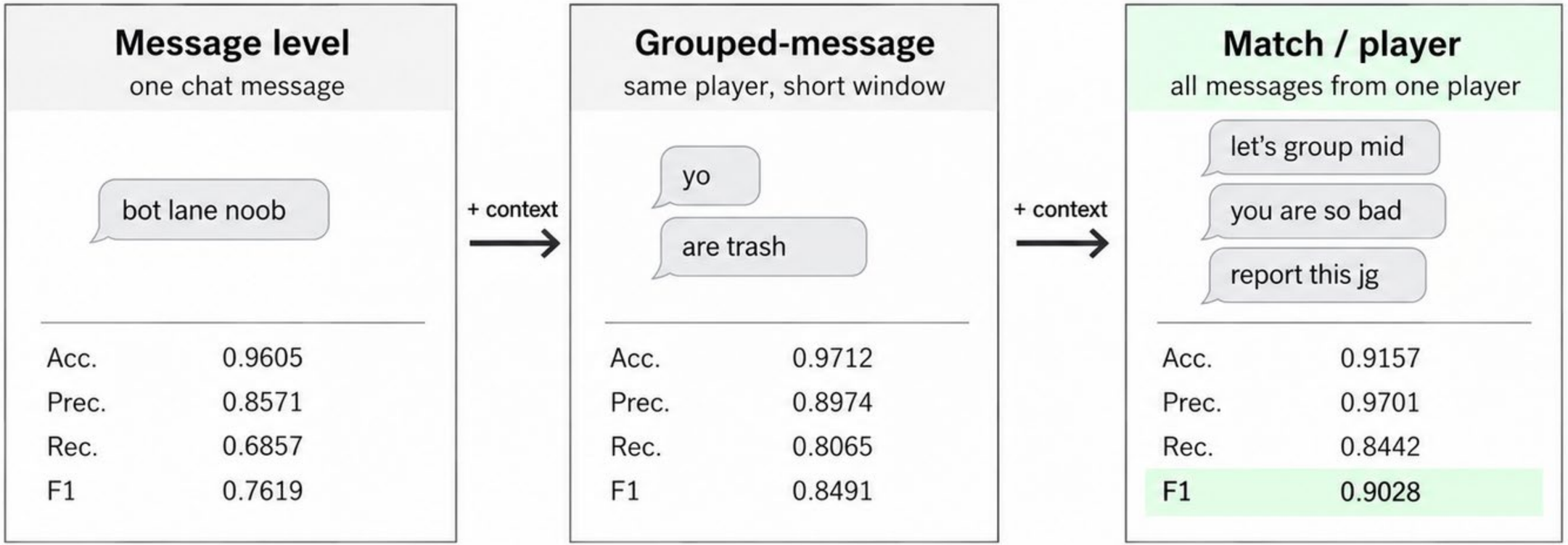
vamos

que haces

MODEL COMPARISON

Model	Acc.	Prec.	Rec.	F1
ProtectAI unbiased toxic RoBERTa (ONNX)	0.9021	0.4659	0.4143	0.4386
Nicholas Kluge toxicity model	0.8708	0.3866	0.6821	0.4935
Junglee BERT toxic comment classification	0.8722	0.3977	0.7500	0.5198
Martin Ha toxic comment model	0.9068	0.4908	0.2857	0.3612
Garak LLM RoBERTa toxicity classifier	0.9074	0.4978	0.3964	0.4414
ChatGPT 4o (June 2025) – P1	0.9008	0.3495	0.9231	0.5070
ChatGPT 4o (June 2025) – P2	0.9178	0.3662	0.6667	0.4727
Llama 3.2	0.7542	0.2451	0.8000	0.3752
Unitary Toxic-BERT (base)	0.9166	0.5401	0.6500	**0.5900**
IGC-BERT (fine-tuned)	**0.9605**	**0.8571**	**0.6857**	**0.7619**

DOES MORE CONTEXT HELP?



OUTSIDE IN-GAME CHAT?

Can the detector work outside in - game chat?



LoL videos

captions →

IGC- BERT

flag lines →

Detected toxic captions
i hear your trash
you're all trash
right you are messed up in the head

- 9 League of Legends video tested
- Captions analysed line - by - line

The detector can be used beyond the original dataset.



- Runs entirely in - browser
- No cloud calls
- Toxic text hidden behind a spoiler tag

545 MB model < 2 min on a heavy webpage

LIMITATIONS

LoL Specific Dataset

Applicability beyond League of Legends is still uncertain.

Prototype Extension

Useful proof - of - concept, not an operational deployment.

Web Demo Is Exploratory

Youtube captions show potential, not a fully validated benchmark.

We show feasibility - not a final moderation system.

KEY TAKEWAY(S)



Gaming toxicity needs gaming context

Generic moderation misses slang, sarcasm and game - specific meaning.



Message - level labels matter

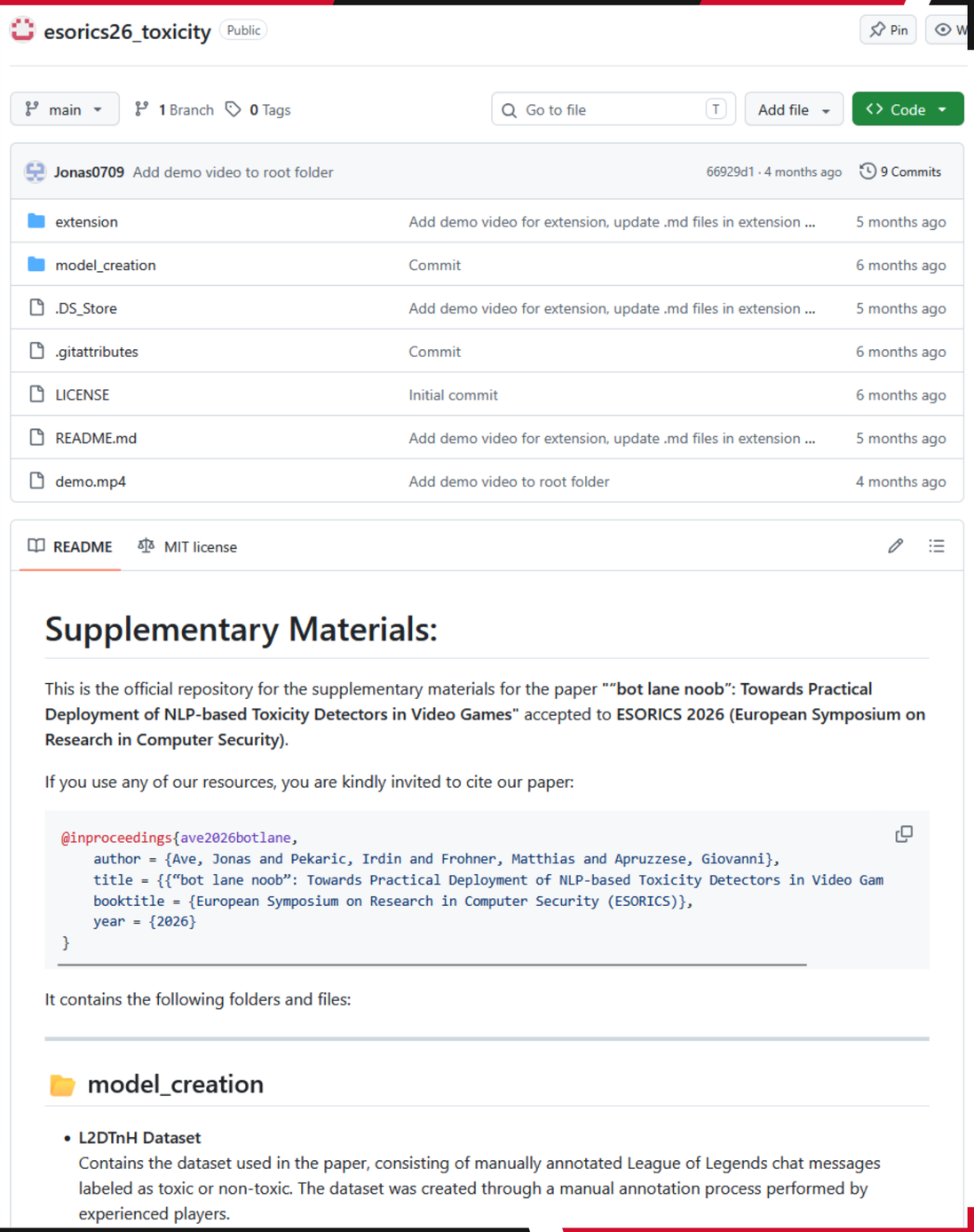
A toxic match is not the same as toxic message.



Open resources make practical detectors possible

Datasets, models and tools can support future work.

Context beats generic moderation.



REPOSITORY

https://github.com/irdin -pekaric/esorics26_toxicity



THANK YOU

FOR YOUR ATTENTION

Irdin Pekaric
irdin.pekaric@uni.li
University of Liechtenstein

ANNOTATOR BACKGROUND

ID	Gaming Exp.	Rank	Hrs/wk	Other Exp.	Msgs
A1	17 yrs	Masters	17	MOBA, FPS, MMOS	5k
A2	20 yrs	Diamond	21	MOBA, FPS, MMOS	12k
A3	9 yrs	Diamond	30	MOBA, SBX, FPS	5k
A4	8 yrs	Gold	28	MOBA, SBX, FPS	5k
A5	6 yrs	Diamond	19	MOBA, SBX, FPS	5k
A6	7 yrs	Bronze	10	MOBA, SBX, FPS	15k
A7	20 yrs	Diamond	20	MOBA, SBX, FPS	5k
A8	9 yrs	Silver	35	MOBA, MMORPG	15k

ANNOTATED CHAT MESSAGE

ID	Message	A1	A2	A3	A4	A5	A6	A7	A8	Label
37	mother fucking noob	1	1	1	1	1	1	1	1	1
83	rammus is such a slut	1	0	0	0	0	1	1	1	1
87	FUCK	1	0	0	0	0	0	0	0	0
114	get fked	1	1	0	0	1	1	1	1	1
123	rammus useless	0	1	0	1	0	1	0	1	1
172	leona you are so bad. you lost this lane gg	1	1	0	0	0	1	0	1	1
203	bot lane noob	1	1	0	0	0	1	0	1	1
213	fucking scumbags blocked the lantern	1	1	0	0	1	1	1	1	1
257	na u suck lucain	1	1	0	0	1	1	1	1	1
315	just uninstall lol	1	1	0	0	0	1	1	1	1