"bot lane noob" Towards Deployment of NLP-based Toxicity Detectors in Video Games

Jonas Ave[1], Irdin Pekaric[1], Matthias Frohner[2], and Giovanni Apruzzese[21]

[1] University of Liechtenstein, Vaduz, Liechtenstein
[2] Reykjavik University, Reykjavik, Iceland

Abstract. Toxicity and harassment are widespread in the video-gaming context. Especially in competitive online multiplayer scenarios, gamers oftentimes send harmful messages to other players (teammates or opponents) whose consequences span from mild annoyance to withdrawal and depression. Abundant prior work tackled these problems, e.g., pointing out the negative effects of toxic interactions. However, few works proposed countermeasures specifically developed and tested on textual messages sent during a match—i.e., when the "harassment" actually occurs. We posit that such a scarcity stems from the lack of high-quality datasets that can be used to devise "automated" detectors based on natural-language processing (NLP) and machine learning (ML), and which can – ideally – mitigate the harm of toxic comments during a gaming session. This work provides a foundation for addressing the problem of toxicity and harassment in video games. First, through a systematic literature review (n=1,039), we provide evidence that only few works proposed ML/NLP-based detectors of toxicity/harassment during live matches. Then, to foster more practical research in this domain, we partner-up with 8 expert League of Legend (LoL) players and create a fine-grained labelled dataset, `L2DTnH`, containing 1.4k toxic and 13.8k non-toxic messages exchanged during LoL matches. We use `L2DTnH` to develop an ML-based detector that we then empirically show outperforms general-purpose and state-of-the-art toxicity detectors reliant on NLP. Finally, to further demonstrate the practicality of our resources, we test our detector on game-related data beyond that included in `L2DTnH`; and we develop a Web-browser extension that flags toxic content in Webpages—without making any query to third-party servers owned by AI companies. We publicly release all of our resources. Our contributions pave the way for more applied research devoted to fighting the spread of toxicity and harassment in video games. **WARNING: Offensive Language**

1 Introduction

Did you know that one out of every three people plays video games [35]? Indeed, the attention towards the video-gaming sector is massive: according to 2025 data, the revenue of the video-gaming industry exceeds $100 billion [74]; video games are the most popular type of mobile application [36]; and "gaming" accounts for 1/10th of the entire internet traffic (ranked third after "video" and "social" [34]). Video games now play a leading role amidst Web-related technologies.

The ecosystem of modern video games is complex. At a high level, creators of video games—spanning from "AAA" studios such as Riot [1] or Blizzard [2], to "indie" studios such as SuperGiant games [3]—make a product for the end users—the players. However, a detailed look reveals that the interactions among these two entities (makers and players) are much more profound. For instance, developers of online multiplayer titles enable players to send messages during a gaming session [79]; gamers also frequently interact via third-party platforms (e.g., Reddit [9] or YouTube [77]), some of which entirely (e.g., Steam [75]) or mostly (e.g., Twitch [80]) depend on game-related content. Unfortunately, such interactions are not always constructive: as abundant reports have pointed out, video games are plagued by the presence of toxicity and harassment [11, 15].

A large body of literature has studied the problem of toxicity and harassment in the video game context [47, 85]. Numerous user studies [18, 45, 69, 78, 88] highlighted the negative consequences (e.g., withdrawal, self doubt, depression) that such messages can have on targeted players. These works typically advocate for the integration of automated mechanism that can deal with the "source" of toxicity/harassment right away—e.g., censorship of the affected messages, or a ban of the offending player. Indeed, recent advances in the machine-learning (ML) and natural-language processing (NLP) domains have the potential to mitigate this problem [24, 31]. Yet, in the research domain, and to the best of our knowledge, only few works (e.g., [55]) proposed and *implemented* solutions, reliant on ML/NLP, to counter toxicity/harassment in the gaming context.

Summary of Contributions. We seek to provide a foundation for the *practical development and deployment* of automated detection mechanisms of game-related toxicity and harassment. We make the following contributions:

- First, we systematically review prior work (n=1,039) and find that only 15 papers proposed ML/NLP-based mechanisms to counter toxicity/harassment (§2). However, such methods always relied on datasets with a limited scope—in terms of openness, source, or labeling granularity. We discuss such limitations.
- Then, to overcome such limitations, we create a novel dataset: `L2DTnH` (§3). `L2DTnH` draws from the well-known Tribunal dataset [86], which reports the chatlogs of League of Legends (LoL) matches that included some (verified) harassment—but without any fine-grained label, making it hardly usable for toxicity detection. So, for `L2DTnH`, we recruited 8 expert video gamers and asked them to use their expertise to annotate the ground truth of *each message* in Tribunal. Overall, `L2DTnH` has 1,398/13,773 toxic/non-toxic messages—making it the largest open-source and game-specific dataset for toxicity detection.
- Next, we show the practicality of `L2DTnH` and use it to fine-tune an ML model (§4). Comparisons with state-of-the-art transformer models for toxicity detection (e.g., Toxic-BERT) reveal our model outperforms all of them.
- Finally, to show the applicability of our tools in the real-world, we tested our model on game data from a different source: captions of YouTube videos (§5). We also developed a browser extension that operates locally and *(i)* does not send any data to remote servers while *(ii)* automatically flagging toxic content in Web pages. We discuss lessons learned from our implementation.

We release all of our resources [4]. We informed Riot Games of our tools.

2 Related Work and Motivation

We outline the concepts and challenges tackled by our paper (§2.1), describe our systematic literature review (§2.2) and finally present the research gap (§2.3).

2.1 Toxicity and Harassment in Video Games

"Toxicity" and "harassment" are pervasive in many online communities [16, 54].

To provide some definitions, "toxicity" denotes behaviors that are harmful towards other individuals, whereas "harassment" indicates targeted and/or repeated instances of toxic behavior, oftentimes resulting in greater harm [51]. For instance, harassment can include severe forms of verbal abuse, or sexism [45]. Yet, a precise distinction between these two terms is ultimately subjective. Hence, in the remainder of this work, we will consider these two terms as synonyms.

The video-game context is particularly prone to toxic behaviors, especially within the communities of competitive online multiplayer games—such as League of Legends [48] (LoL). This specific types of games, occasionally referred to as "esports" and counting almost 1 billion users [73], presents peculiarities that make identification of toxic behaviors through NLP more challenging than in other domains. For instance, terms such as "noob" or "uninstall" can be used to mock other players, but are meaningless outside a gaming context. At the same time, specific games have their own toxic jargon (e.g., the statement "Leona is botting" only has toxic implications within LoL). Such specificity makes general-purpose solutions against toxic behaviors (e.g., censoring of bad words) not very effective in the gaming ecosystem, thereby calling for of ad-hoc mitigations.

2.2 Systematic Literature Review

Given the challenges of fighting toxicity and harassment in the video-gaming context, we wondered: "what prior works proposed automated techniques, based on ML/NLP methods, to detect toxicity/harassment in the gaming context?" We tackle such a research question (RQ) via a systematic literature review (SLR).

Paper Collection. Our SLR follows established PRISMA guidelines [62]. The entire procedure was conducted by two authors who interacted and validated each-other's findings. First, between Dec. 2024 and March 2025, we queried four popular databases of peer-reviewed literature (ACM DigitalLibrary, SpringerLink, IEEE Xplore, Elsevier ScienceDirect) for papers matching the queries ("game/player/esport" $\wedge$ "ML/AI/NLP" $\wedge$ "toxicity/harassment") published since 2014; we perform our searches between Dec. 2024 and March 2025. Such a search yielded a 1,039 papers. The exact search terms are in our repository [4].

Screening. We then manually checked the metadata (title and abstracts) of these papers, removing those that were clearly outside of our scope (e.g., a user study with no solution [65]). Such a process led to excluding 989 papers. The remaining 50 papers were then analysed in their entirety. Such an analysis, led to the removal of 35 papers (e.g., no application of AI [72], or applications with no

relevance to NLP [23]). Hence, out of 1,039 papers, only 15 (<2%) proposed/implemented some automated mechanisms to address toxicity/harassment.

Findings. Altogether, these 15 papers have provable limitations. Some focus on "emotes" [46] or on detecting toxic content not during live gameplay (e.g., Twitch chats [30, 56], YouTube [61], reviews [81], online forums [82]). Some do not provide details or do not share the dataset used to test a given hypothesis [38, 68, 76]. Others [26, 70] do not focus on the English language (i.e., the de-facto language of multiplayer online games [55]). The most relevant works are: [19, 55, 58, 60]. However, [55] only focus on verbal abuse, overlooking other forms of toxic behavior such as trolling; whereas [58] combine various data sources, including chat logs processed by ML, but focus on a game (World of Tanks) that has long since stopped being popular (e.g., 4k current players in October 2025 [14]). Finally, [19] and [60] focus on LoL (still extremely popular [13]) and use the Tribunal dataset which provides aggregated data of entire matches.

2.3 Research Gap and Problem Statement

Our SLR highlights that *(i)* despite abundant research interest in game-specific toxicity/harassment, *(ii)* only few works specifically proposed ML/NLP methods to address this problem within the game itself, and *(iii)* previously-proposed contributions have some limitations from a practical viewpoint—as also evidenced by the fact that toxicity/harassment are still an open problem [11].

We argue, echoing the opening statement of [58], that one of the reasons why ML/NLP methods have not seen more applications in this context is due to lack of high-quality data. Such a lack prevents both *(a)* development of ML/NLP detectors, and *(b)* assessment of any sort of mitigation (not necessarily reliant on ML/NLP). By "high-quality data" we specifically refer to a large-scale dataset provided with granular ground truth that reflects the specific types of toxic behaviors and which integrates game-specific knowledge.

Therefore, our primary objective is contribute with a dataset that fulfills the aforementioned requirements. We acknowledge that there are ways to use ML that do not encompass NLP techniques (e.g., analysis of gameplay elements [23] or audio features [66]). These approaches are valid, but orthogonal to our goal.

Positive Light. We do not seek to invalidate prior research. Moreover, due to lack of public code, we cannot reproduce prior contributions. We are merely elucidating factual shortcomings that motivated us to pursue our research objective. Our open-source [4] contributions are rooted on prior work.

3 Our Proposed `L2DTnH` Dataset

We describe our proposed `L2DTnH`, short for "`L`oL-based `L`abeled `D`ataset of `T`oxicity and `H`arassment." We first outline the origin and characteristics of the source data (§3.1), then present the labeling procedure (§3.2), and finally provide quantitative metrics (§3.3). An overview of our methodology is shown in Fig. 1.

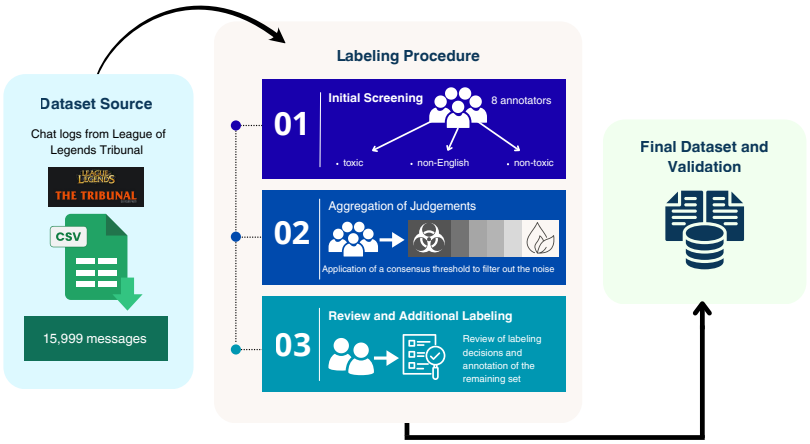

Fig. 1: Overview of the creation process of L2DTnH.

3.1 Context and Challenges

Our research builds upon the LoL "Tribunal" chatlogs dataset [86]. The dataset was created by RIOT Games' "Tribunal" moderation system [12], an initiative where experienced players collectively review matches wherein an player had been reported to exhibit toxic behavior—with the purpose of verifying if such claims are true. Each case in the Tribunal dataset contains complete in-game chatlogs and metadata (e.g., player roles, champions, team), including the decision obtained by majority voting among reviewers.

The Tribunal dataset is valuable because it contains large-scale (>1M chatlogs) real-world communications between players in a highly competitive online environment that have been verified as being toxic. Unlike social media platforms or forums, LoL chat contains fast-paced exchanges, fragmented utterances, as well as strong emotional reactions. The dataset thus contains player interactions that include sarcasm, slang, and game-specific abbreviations that are rarely found in general-purpose toxicity datasets (e.g., [84]).

Despite such advantages, the Tribunal dataset has a notable shortcoming: **annotations are provided only at the match level**. Indeed, if a player in a match is deemed "toxic", then the entire match is considered as "toxic". This means that all messages from a toxic match are considered toxic—even neutral or supportive ones (e.g., messages such as "w8" or "ok" are treated as "offensive language" simply because they appear in a toxic match). In other words, the information provided in the Tribunal dataset does not allow to identify the specific toxic messages. Such a limitation hence inhibits carrying out message-level classification of toxic content.

3.2 Fine-grained Annotation

We carried out our annotation to converts the original "match-level" Tribunal dataset into a "message-level" resource, i.e., our proposed L2DTnH dataset.

Table 1: Annotator background and labeling contribution.

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

Annotator selection We recruited eight annotators who were all long-term LoL players with 6–20 years of experience in the gaming domain. Table 1 summarizes their backgrounds, including average LoL weekly playtime and in-game rank. Prior familiarity with the specific game's linguistic culture was a key selection criterion, as gaming toxicity is frequently expressed through linguistic cues such as sarcasm ("nice ult bro"), abbreviated insults ("ez"), or creative misspellings ("n0000b"). Annotators unfamiliar with these conventions can misinterpret intent, introducing substantial bias.

Labeling procedure Labeling each message of the over 1M chatlogs within the Tribunal dataset is not humanly possible, especially because determining whether any given message is toxic or not is ultimately subjective [22, 52].[3] So we devised a best-effort strategy founded on consistency, organized in three steps.

i) First, each annotator was assigned the same set of 5,000 messages,[4] and was tasked to independently assign the label "toxic,"[5] "non-toxic,"[6] or "non-English". Annotators were explicitly instructed to base judgments solely on linguistic cues—potentially accounting for quick repeated messages sent by the same player in a short timeframe (e.g., "yo", "are", "trash").

ii) Then, to handle subjectivity, we aggregated the judgments by applying a consensus threshold: a message was deemed as "toxic" if at least two annotators labeled it as such (see Table 2).[7] Such a procedure enabled to filter out noise, while ensuring that a "toxic" message could be truly perceived as toxic.

iii) Finally, three annotators reviewed the labeling decisions and proceeded to annotate an additional set of 10,999 messages (one only participated in 7k). For this step, disagreements or borderline cases were discussed jointly until a final consensus was reached.

Our annotation hence increased the number of labeled instances by 240% without demanding additional human-working hours by five of the eight annotators

[3] Similar challenges in creating labeled datasets are previously reported (e.g., [67]).

[4] We took the first 5k messages in Tribunal, which can be considered as random sampling since the distribution of toxicity across Tribunal can be assumed as random.

[5] E.g., insults, harassment, offensive tone, or otherwise suspicious content.

[6] In principle, ambiguous cases such as sarcasm or playful teasing are toxic only if a player would reasonably perceive them as offensive within a competitive match.

[7] This threshold was determined after continuous assessments, and we occasionally even inquired the specific annotator(s). Ultimately, toxicity is subjective and if two people agree a message has toxic potential, then it should be deemed as such.

Table 2: Examples of annotated chat messages with decisions.

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

beyond that of the first 5k messages—which have been used to "align" the mindset of the last three annotators with that of entire set of annotators.

Validation metrics To objectively validate our annotation strategy, we computed well-known statistics after the second step. The Fleiss' $\kappa = 0.62$, indicates substantial agreement among the eight annotators [37]. Overall, 4,704 messages (out of 5,000) were unanimously marked as "toxic", "non-toxic", or "non-English" demonstrating strong consensus on highly-abusive expressions. Hence, the labeling was consistent across our sample. Finally, $\approx 1.9\%$ messages were flagged as toxic only by a single annotator, and hence resolved as non-toxic.

3.3 Quantitative Analysis

At the end of our annotation procedures, we obtained a set of 15,999 labeled messages, representing the chatlogs of ≈100 different LoL matches containing some (verified) instances of toxicity.

Specifically, out of 15,999 messages, 1,398 (8.74%) are "toxic", 13,773 (86.09%) are "non-toxic" and 828 (5.17%) are "non-English". Across the 15,171 English messages, the average length (in characters) is 12.5 (std=11.53).

We can hence approximate the overall "cost" of our labeling efforts. A conservative assumption is that each message required 5 seconds to be labeled (which includes: reading the message, and potential surrounding messages; determining its ground truth; and physically assigning the label). Hence, the first step of our labeling procedure required: 5 seconds × 8 annotators × 5,000 messages = 200k seconds ∼ 56 hours; whereas the third step required: $5 \times (2 \times 10{,}999 + 7{,}000) =$ 145k seconds ∼ 40 hours. Taking into account discussion and reviewing, we can hence estimate that `L2DTnH` required ≈100 human-working hours to be created.

4 Using `L2DTnH` in Practice: Empirical Tests

It is factual that our `L2DTnH` represents a valid testing ground for toxicity detectors. However, a question arises: "Does `L2DTnH` improve existing ML- and NLP-based toxicity detectors?" We answer this question with an original assessment of various models reliant on pre-trained transformers. We first describe the experimental setup (§4.1), then present the results (§4.2); finally, we examine the effects of using `L2DTnH` at different levels of message grouping (§4.3).

4.1 Testbed and Model Development

An investigation of the current landscape of pre-trained transformers reveals that Toxic-BERT [40] is among the best general-purpose models for toxicity detection, so we consider this as a baseline.

To tailor this model to the gaming domain, we fine-tune it on our `L2DTnH`. Thus, we take the 15,171 English messages in `L2DTnH` and use stratified sampling to partition them into a training and testing set by applying an 80:20 split. Fine-tuning was implemented using the Hugging Face Transformers and PyTorch frameworks. The model's original six-class classification head was replaced by a binary classification layer. Input sequences were tokenized with the BERT tokenizer (maximum sequence length = 192 tokens). Training was performed for four epochs with a batch size of 64, using the AdamW optimizer (learning rate $= 2e-5$) and a linear learning-rate scheduler. Early stopping with a patience of two epochs was applied to prevent overfitting. We set the CrossEntropyLoss as the objective function. Implementation details are in our repository [4].

At the end of this fine-tuning procedure, we obtain a new model which we denote IGC-BERT (Inappropriate Game Chat-BERT).

4.2 Evaluation Results and Comparison

For a broad a comprehensive assessment, we compare our IGC-BERT against state-of-the-art and publicly-accessible models.

Specifically, we consider: *protectai_unbiased_toxic_roberta* [5], *nicholas_kluge_toxicity_model* [27], *junglelee_bert_toxic_comment_ classification* [6], *martin_ha_toxic_comment_model* [7], *garak_llm_ roberta_toxicity_classifier* [53], and the baseline *unitary_toxic_bert* (as an ablation study). All such models are adapted to a binary classification setting. The assessment is always done on the same test set (i.e., 20% of `L2DTnH`). The entire codebase is in our repository [4].

We also considered two large-language models: first, **ChatGPT** and, specifically GPT-4o (the June 2025 version). We used its API to make our tests, which required us to purchase API queries and forced us to carry out a smaller evaluation on a subset of 667 non-toxic and 40 toxic messages (≈1/3 of our test set). Given its multi-purpose focus, we devised two one-shot prompts for assessing ChatGPT: P1="You are a classifier that detects toxic behavior in gaming chats. Classify the following message as either 'toxic' or 'non-toxic'. The message may contain slang, sarcasm, abbreviations, or profanity. Your response must be only: toxic OR nontoxic." and P2="You are a classifier that detects if a message in gaming chats is inappropriate. Classify the following message as either 'inappropriate' or 'non-inappropriate'. The message may contain slang, sarcasm, abbreviations, or profanity. Your response must be only: inappropriate OR non-inappropriate." Hence, we made ≈1,400 queries to OpenAI's API, and recorded the answers from GPT-4o. Then, we considered another LLM: the (free) **Llama 3.2**, using a variant of P1 (reported in our repo [4]) which we adjusted for the characteristics of Llama 3.2.

The results of our assessment are shown in Table 3, reporting Accuracy, Precision, Recall, and F1-score (we consider a "positive" as a toxic message). For

Table 3: Comparative performance of state-of-the-art toxicity detection models on the test portion (20%) of our L2DTnH dataset. For cost reasons ChatGPT was tested on a subset of our test set.

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
junglelee_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**	**8571**	**0.6857**	**0.7619**

completeness, the test set has 2,755 non-toxic and 280 toxic messages. We appreciate that fine-tuning on our L2DTnH yielded statistically significant improvements (validated with a t-test at $p > .05$) over the baseline model, with improvements of nearly 20 absolute percentage points in the F1-score, and the false positives decreased from 137 to 32. We investigated these results: the baseline model classified neutral or sarcastic expressions as toxic, due to lack of adaptation to the game/LoL-specific domain. ChatGPT 4o and LLama 3.2 do not seem to be very effective, as indicated by underwhelming precision (always below 0.4).

Takeaway. Fine-tuning on our proposed L2DTnH leads to statistically-significant ($p < .05$) improvements: our IGC-BERT model outperforms existing general-purpose toxicity detectors (thanks to our labeled dataset).

4.3 Examining different aggregation techniques

We study the effects of using our proposed L2DTnH by accounting for different ways to aggregate the messages contained therein.

Rationale. Toxic communication in multiplayer games rarely occurs in isolation. Short, fragmented chat messages accumulate into larger sequences that reflect emotional escalation or interpersonal conflict. To capture these contextual effects, we scrutinize our fine-tuned IGC-BERT model at three hierarchical levels of granularity: *(i) message level*—single chat entries evaluated independently; *(ii) grouped-message level*—consecutive messages from the same player in a short time frame combined into a single utterance; *(iii) match level*—aggregation of all chat messages produced by one player during an entire game.

Setup. For a fair and consistent evaluation, we followed a slightly different workflow than that presented in §4.1, which assumed a random split. This is because we cannot create chains of "grouped messages" (from the test set) if each message is drawn randomly. So, for this assessment, we create the training set by considering all messages exchanged in 81 games (i.e.,≈80% of all matches in L2DTnH) and the test set by considering all matches of the remaining 18 games (i.e., ≈20% of all matches in L2DTnH). We hence used the training set to create another fine-tuned variant of our IGC-BERT.

Results. We report the results in Table 4. Specifically, for the message-level results, we simply tested on the (new) test set (in the same fashion as in §4.2) and

Table 4: Performance of IGC-BERT across different contextual granularities of message aggregation (note: the IGC-BERT model here follows a different fine-tuning process than that in Table 3)

Evaluation Level	Acc.	Prec.	Rec.	F1
Message level	0.9605	0.8571	0.6857	0.7619
Grouped-message level	0.9712	0.8974	0.8065	0.8491
Match level	0.9157	0.9701	0.8442	0.9028

we see that the performance aligns with that shown in the last row of Table 3: small differences in the recall can be explained with a different distribution of toxic messages in these games, which are not captured during the fine-tuning of this variant of IGC-BERT. For the Grouped-message level results, we fairly aggregate messages of the same player within a short timeframe (e.g., 10s) and input such "longer" messages to the model: we see a substantial increase in the performance (i.e., the recall increases by 12 absolute percentage points) because the model can better recognize toxic content which would be otherwise missed by few-word messages. For the Match-level results, we aggregate all messages of the same player together and submit it to the model. Such an evaluation yields a very high precision with almost no false-positives (only 2 players have been falsely flagged as being toxic), because the model is always able to pinpoint at least some instances of toxicity within the (very long) string received as input.

Takeaway. By fine-tuning BERT on the individually-labeled messages in `L2DTnH`, one can develop a model that precisely detects toxic players in a match.

5 Beyond the dataset: using our tools on the Web

Insofar, we have evaluated existing (and new) models on our proposed `L2DTnH` dataset. Here, we further demonstrate the practical value of our contributions by exploring "unknown" game-related contexts. First, we test our IGC-BERT model on LoL videos on YouTube (§5.1). Then, we integrate IGC-BERT in a custom-made browser extension (§5.2). Finally, we compare `L2DTnH` against other datasets for toxicity detection (§5.3).

5.1 Testing IGC-BERT on LoL YouTube Videos

We find it instructive to evaluate the performance of IGC-BERT on YouTube videos entailing LoL-related content.

Disclaimer. To avoid fingerpointing exercises, we deliberately chose videos that are sarcastic and/or whose creators are well-aware that the video includes some form of toxicity. We merely seek to gauge if our model can detect possible toxic instances in captions—which are semantically different than chatlogs.

Why is it relevant? We designed our `L2DTnH` to capture instances of toxicity/harassment during a LoL match. However, similar instances can occur also outside of a match. For instance, YouTube is a popular resource for LoL-related content: players can upload videos of their games, and share them on the Web.

Unfortunately, some videos can include toxic content, e.g., imprecations of the creator after being defeated by an enemy, or "trash talking" a certain opponent with game-specific jargon. Even though such occurrences may not reach the actual player, downstream viewers can be hurt by such content, or can find it disrespectful—which can lead to, e.g., the video being reported or taken down. Therefore, content creators can use our resources to automatically analyse their videos *before* uploading them, thereby preventing harmful consequences.

Methodology. Our goal is simple: assessing the extent to which our IGC-BERT model can detect toxic content in LoL videos uploaded on YouTube. We note that our IGC-BERT model expects *textual data* as input. So, to enable our model to analyse YouTube videos, we download the *captions* (automatically generated by YouTube) and submit them, line-by-line, to our IGC-BERT model. However, and unfortunately, we are not aware of any publicly-available dataset of YouTube-videos' captions that contain verified toxic instances of LoL-related gameplay. Hence, our assessment is just a proof-of-concept experiment, meant to demonstrate an ancillary (but practical) application of our resources to address toxicity/harassment in the gaming context. Therefore, we identify a total of 9 YouTube videos which we expect may contain some toxic content (we found them by submitting the query "lol toxicity" on YouTube), retrieved their captions, submitting them to our IGC-BERT model (developed in §4.1) and analysed if our model could classify any line of the captions as "toxic", thereby indicating that the audio of the YouTube video has toxic content.

Results. We report below the 9 videos (clicking on the title leads to the video on YouTube) and an exemplary "toxic" captions line (according to IGC-BERT).

1. Best Trolls of 2021 League of Legends "i hear your trash"
2. League of Legends is somehow getting MORE Toxic "this bondage boy sex ring jail escap who"
3. Toxic ADC flames me... but he doesn't know I'm the Rank 1 Senna "completely worthless baby raging ADC but"
4. Why League Is The MOST TOXIC Game of All Time "How can you all be this fed and cry "
5. EUW IS THE MOST TOXIC SERVER IN LEAGUE OF LEGENDS HISTORY "vagar you little bell pepper you even"
6. I wanted to quit... #10 "right you are messed up in the head"
7. The most toxic player in League #8 "cannon no a jungler has mental"
8. Why League of Legends is SO TOXIC _ League of Legends "game because you're all trash stay"
9. Why League of Legends' Design Encourages Toxicity _ Design Delve "some dude saying YOUR MUM, FAT BUM, WIDE TUM"

Such a proof-of-concept experiment indicates that our IGC-BERT model can be applied to detect toxicity/harassment in LoL-related YouTube videos. Such a mechanism can not only be used by content creators to preemptively issue warnings to their audience, or reconsider uploading their video, but can also be integrated in video-sharing platforms (not limited to YouTube) to inform their users (creators or watchers) that a given video may contain toxic content. N.b.: the added value of IGC-BERT is that it flags toxic content pertaining to game-specific jargon—which is seldom captured by general-purpose toxicity detectors.

5.2 Development of a Browser Extension

We developed a browser extension that uses our IGC-BERT model to automatically censor certain elements of Web pages that are flagged as toxic. We first

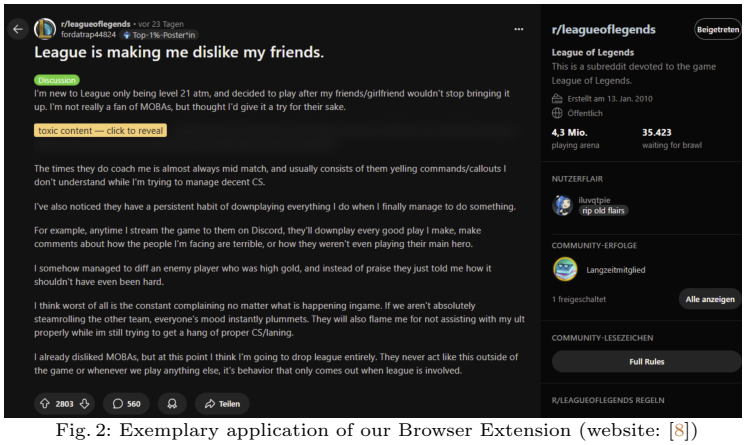

Fig. 2: Exemplary application of our Browser Extension (website: [8])

explain why we did so, then outline the design objectives and the implementation choices and presenting the technical requirements of our extension.

Motivation Using a browser extension to automatically detect toxic content is not new. However, in October 2025, we carried out a literature review by querying Google Scholar with the strings ("browser extension" $\wedge$ "harassment/toxicity"), considering the first 100 results of each query. We found only 8 peer-reviewed papers that truly attempted practically implement a browser extension for toxicity detection. Unfortunately, most of these (i.e., [20,21,29,64,71]) do not release their source code, preventing reproducibility. Others (e.g., [43,50]) do not focus on toxicity/harassment, or not on the English language (e.g., [71]) or not on the gaming domain (e.g., [44]). Such a landscape motivated us to develop our extension and share our lessons learned in this paper.

Objectives Our extension has one purpose: analyse web pages, and proactively censor any content that is toxic, preventing the end-user from reading harmful messages. We envisioned such a process as a two-step approach: first, when a user lands on a webpage, the HTML is sent to the extension; then, the extension analyses the HTML via our IGC-BERT model and, if it detects toxic content, it conceals it under a "spoiler" tag that can be removed at user discretion (see Fig. 2). The extension works entirely in-browser: we explicitly forbid executing any sort of query to remote servers with the purpose of, e.g., having powerful ML-based models analyse the Web pages browsed by the user and flag potential toxic elements (as done, e.g., in [21], which uses Perplexity AI)).

Implementation We explain the technical implementation of our extension.

- All inference is performed locally in the browser (no cloud connections, usable offline). For implementation, `web_accessible_resources` are used for `vendor/*`, `models/`, `ml/`. Dynamic import is performed using `import(chrome.runtime.getURL(...))` in the content script and a local interface (`initIfNeeded()`, `runDetectorBatch()`) in `ml/inference.js`. As a result, content does not leave the page, and the extension works without an internet connection.
- A quantised ONNX model [33] was used for the model implementation. This choice was made due to the reduced memory requirements and shorter loading

times, as well as faster inference compared to non-quantised models. Indeed, impossibility to rely on external API calls meant that all computing had to be done locally, so it was crucial to minimise resource expenditure.
- The inference of content in a web page was batch processed to reduce hang-ups. The implementation was carried out with `runDetectorBatch()` using the configuration parameters of our custom IGC-BERT model. The result was a more stable user interface, fewer system failures and consistent progress.
- An issue we encountered was handling of abortions, which triggered duplicate scans. Such occurrences were frequent when changing tabs quickly changing webpages. So, we used a flag (`aborted=true`) that would trigger an early stopping of the scanning of the webpage if necessary.

Since the scan occurs when the page loads, the extension is not designed to work on content that loads progressively (e.g., the news-feed of popular online social networks). Dynamic scanning substantially increases the computational load.

Technical requirements Our final version of the extension (for which we provide its interface in Fig. 3) requires 545MB of free space, most of which (95%) are for storing the model. We tested the extension on a "heavy" webpage with a lot of text (i.e., [8]). Earlier version of the extensions required 4GB of RAM and took more than 5m to process everything and resulted in crashing the browser; the most recent version takes less than 2m to do so (with no crashes). We monitored the memory utilization during the processing of the testing webpage: in "idle" (i.e., with the default splash screen), the browser absorbs 520MB, which raises to 780MB after landing on [8]; during the 2m of the scan, the used memory increases to 1,400MB; after the scan ends, the used RAM drops to 1,300MB. These tests have been done on a machine equipped with an AMD Ryzen 9 7950 (the quantised IGC-BERT model runs on CPU) and 32GB of RAM.

Takeaway. Development of a browser extension that integrates our model is possible, but processing heavy webpages may require long loading times. By offloading the computation to third-party servers, such analysis would be faster—but such responsiveness would sacrifice privacy. We provide in our repo [4] a 30s demo of our extension. Moreover, we show in Fig. 2 a screenshot of our extension, which blocked some toxic content in a LoL subreddit.

5.3 Other Datasets for Toxicity Detection

Our `L2DTnH` is not the only dataset usable for game-related toxicity detection. Let us fairly compare our proposed dataset with those used by prior work.

To this end, we consider the datasets we found mentioned in the papers we analysed during our SLR (in §2.2). For each of these datasets, we consider: whether they are publicly accessible; their number of samples; the data source (e.g., forum posts, or chat messages); the granularity; the game/platform from which they are taken. We report the results of our analysis in Table 5.

Most prior works considered datasets that are not publicly available. Three works (i.e., [19, 49, 60]) considered *subsets* of Tribunal which are not released, and the artifacts are not accessible (as of March 2026) preventing reproducibility.

Among those for which the dataset is publicly available, we mention: "Corpus-Twitch-Videogames" and the custom one in [30], containing chat messages sent on Twitch (i.e., not during a match); and the Cyberbullying dataset, which refers to LoL and WoW, but contains forum messages. Hence, to the best of our knowledge, our proposed `L2DTnH` is the largest publicly-available dataset containing fine-grained and entirely manually-labeled messages[8] sent during LoL in-game matches, and which are valid for toxicity detection.

We also mention the existence of other datasets for toxicity detection (e.g., [28, 39, 41]), but not meant for the video-game context.

6 Discussion

We outline the limitations of our research (§6.1), discuss lessons learned (§6.2), and make some ethical considerations (§6.3).

6.1 Limitations, Disclaimers, and Threats to Validity

Our overarching goal is to spearhead development of practical solutions for countering toxicity/harassment in the gaming context—which are needed today.

We acknowledge that our SLR presents the limitations of querying databases. For instance, our search results did not include a related work [87], which surprisingly was not returned by the ACM DL (despite matching our search terms). Still, the paper [87] is complementary to ours, as it focuses on different games (For Honor, and Rainbow Six Siege) and leverages proprietary data.

The creation of our `L2DTnH` (in §3) was done by finding a trade-off between labeling quality and dataset size. The original Tribunal dataset has millions of messages and labeling all of them is beyond the capabilities of a single research unit. Still, our results (in §4 and in §5) show improvement over the baseline and utility both within our dataset and in unknown (but related) domains. Importantly: we do not claim that `L2DTnH` can be used for toxicity detectors of *any* game: `L2DTnH` contains data pertaining to LoL, so its applicability beyond LoL is uncertain. Therefore, we do not find any threat to the validity of our conclusions concerning the utility of our proposed `L2DTnH` for its intended purpose.

The experiments done in unknown domains are a proof of concept, and our browser extension is just a prototype (§5). We do not claim that our resources can be deployed in operational systems. Also, we underscore that the superior performance of IGC-BERT over general-purpose models (see Table 3) is due to fine-tuning on our proposed `L2DTnH`, and not due to methodological novelty.

6.2 Lessons Learned and Implications

First, our SLR showed that, despite a great interest in this topic, few papers propose and implement (and share) solutions to address the problem of toxicity and harassment in video games. Such a finding should serve as a call for action.

[8] E.g., the recent GameTox dataset [59], which is unrelated to LoL and was not captured in our SLR, is larger but labeling relied mostly on LLMs which are error-prone.

Table 5: Datasets Used in Game-Related Toxicity Research

Dataset	Access	Gran.	Size	Source	Game/Platform	Used in
Corpus-Twitch	✓	message	2.6k	chat msgs.	Twitch	[56]
WotReplays	✗	message	15k	replays	WOT	[58]
Custom Dataset	✗	phrase	765k	chat msgs.	RO, DOTA	[26]
Custom Dataset	✓	message	100k	chat msgs.	Twitch	[30]
Dotalicious	✗	word	7M	chat msgs.	DOTA	[55]
Cyberbullying dataset	✓	comment	34k	forum	LoL, WoW	[82]
Tribunal dataset	✗	report	11M	chat msgs.	LoL	[49]
RIOT dataset	✗	thread	89	forum	LoL	[68]
Tribunal dataset	✗	report	2M	chat msgs.	LoL	[60]
Tribunal dataset	✗	report	11M	chat msgs.	LoL	[19]
CONDA	✗	message	50k	chat msgs.	DOTA	[83]
For honor dataset	✗	chat	1800	chat msgs.	For Honor	[23]

Second, our open-source `L2DTnH` serves as a foundation for future research in this domain. Thanks to `L2DTnH`, it is possible to develop, or test, novel countermeasures. Indeed, when we begun this study, we were surprised to find a lack of publicly-available resources that are *(i)* game-specific and *(ii)* labeled at the message level. Such a lack supports our hypothesis that the panorama of practical countermeasures to toxicity/harassment is limited from a research viewpoint.

Third, and anecdotally, while testing IGC-BERT on YouTube (in §5.1) we attempted to test it also on videos that, in our opinion, did not have any sort of toxic behavior. So, we downloaded the captions taken from some videos meant for kids (Specifically, we used: Caillou at the Restaurant) We were surprised when we noticed that our model found instances of toxicity. Upon checking, we found that the following lines had been flagged as toxic: "what's a donkey doing here" and "your toy when your dad gets back". We find such a result fascinating: such lines, within a kids' show, are clearly free of any toxic content; however, by hypothesizing that such text is sent during a LoL match, the meaning substantially changes. Our takeaway is that such findings demonstrate the challenges, but also the need, of developing context-specific detectors of toxicity. This result further confirms our claim: ultimately, **our contributions are designed for LoL. It is unrealistic to expect effectiveness in other games.**[9] Such an objective can, however, be pursued via ensembles (e.g., developing multiple game-specific detectors, and using the one specific for the given context).

Lastly, integrating our IGC-BERT model in a browser extension working locally presents tradeoffs. These can be mitigated by using, e.g., model distillation [42] to produce a "smaller" model that can work faster, but at the expense of detection performance. Such engineering endeavours are beyond our scope.

6.3 Ethical Considerations

When we carried out our research, our institutions did not have any formal IRB process. However, we performed our study according to best practices [17].

[9] In our repository [4], we have evaluated our considered models also on a dataset of 3k Dota2 messages (as also done in [32]) and on the 1k messages in YouToxic (as also done in [57]). These experiments are orthogonal to our main objective. However, our findings confirm our hypothesis: our IGC-BERT does not work so well in contexts different from LoL. We consider this empirical finding crucial for future work.

For our L2DTnH dataset, we relied on the expertise of 8 veteran LoL players. Their participation in such an effort was voluntary, and they willingly agreed to contribute to this study. Participants were aware that their actions would be used to produce a dataset used for research purposes, which was planned to be publicly shared afterwards. We did not collect any sensitive or personally-identifiable information about our participants [10, 25]. To preserve the anonymity of our participants, we cannot disclose additional details about them.

We do not envision any potential negative outcome from our study. We warned that our resources are prototypes, and should not be used to, e.g., claim that any given user is exhibiting toxic behavior without any additional form of validation. In contrast, openly releasing our tools outweighs any risk, since it is the best way to address the widespread problem of toxicity and harassment in video games: sharing our resources enables future work to build upon our findings, fostering more (and much needed) research in this domain.

7 Conclusions and Future Work

Despite the benefits that video-games bring to our lives, thousands of players still report being victim of targeted toxicity/harassment. This is particularly true in competitive multiplayer titles, where the goal is not always limited to "playing for fun", leading to players forgetting that there are human beings on the other side of the screen, who can be severely hurt by reading certain messages.

We tackled this problem and found that, unfortunately, there are no datasets with labeled instances of toxic messages exchanged during a match. So, we created a novel dataset, L2DTnH, thanks to the contributions of 8 veteran players of the ever-popular LoL game. We used L2DTnH to carry out a number of experiments, such as developing an ML-based model that outperforms general-purpose toxicity detectors, testing the model on YouTube videos, and integrating the model on a browser extension. Altogether, our resources should inspire future work focused on practical solutions to the problem of toxicity and harassment in video games. We reached out to RIOT to inform them of our solutions.

We identify three avenues for future work. First, expanding our L2DTnH dataset by labeling additional instances taken by the Tribunal dataset, or testing our IGC-BERT model on messages in the Tribunal dataset not included in L2DTnH. Second, the development of datasets (or models) focused on different games than LoL. Third, the integration of our resources with orthogonal detection techniques, such as those based on gameplay/behavioral elements (e.g., correlating chatlogs with "repeated pings" or other types of griefing behavior [63]), that can further augment the identification of toxic behavior during a match.

Acknowledgments. We would like to thank the anonymous ESORICS'26 reviewers for the great feedback. Parts of this research has been funded by Hilti.

References

1. https://www.riotgames.com/en

2. https://www.blizzard.com/en-us/
3. https://www.supergiantgames.com/
4. https://github.com/irdin-pekaric/esorics26_toxicity/
5. https://huggingface.co/protectai/unbiased-toxic-roberta-onnx
6. https://huggingface.co/JungleLee/bert-toxic-comment-classification
7. https://huggingface.co/martin-ha/toxic-comment-model
8. Accessed on October 7th, 2025, https://www.reddit.com/r/leagueoflegends/comments/1ngnf9b/league_is_making_me_dislike_my_friends/
9. Reddit gaming, https://www.reddit.com/r/gaming/
10. Sensitive personal data. https://home.treasury.gov/taxonomy/term/7651
11. Forms of harassment in online video games in the united states in 2023. Tech. rep., Statista (2023), https://www.statista.com/statistics/1133182/harassment-online-video-games/
12. The tribunal. Accessed on October 6th, 2025 (2025), https://leagueoflegends.fandom.com/wiki/The_Tribunal
13. ActivePlayer: League of legends live player count and statistics. Accessed on October 6th, 2025 (2025), https://activeplayer.io/league-of-legends/
14. ActivePlayer: World of tanks (wot) live player count and game statistics. Accessed on October 6th, 2025 (2025), https://activeplayer.io/world-of-tanks/
15. ADL: Free to play? hate, harassment and positive social experience in online games 2020. Accessed: 6 Oct. 2025 (2020), https://adl.org/resources/report/free-play-hate-harassment-and-positive-social-experience-online-games-2020
16. Aroyo, L., Dixon, L., Thain, N., Redfield, O., Rosen, R.: Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In: Companion Proceedings of TheWebConf (2019)
17. Bailey, M., Dittrich, D., Kenneally, E., Maughan, D.: The Menlo report. IEEE Security & Privacy (2012)
18. Beres, N.A., Frommel, J., Reid, E., Mandryk, R.L., Klarkowski, M.: Don't you know that you're toxic: Normalization of toxicity in online gaming. In: CHI (2021)
19. Blackburn, J., Kwak, H.: Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In: WWW (2014)
20. Bonthu, B., Abhay, P., Gottipati, L.S., Vamsi, G.K.: Civilitycheck: An integrated natural language processing and machine learning framework to detect hateful and offensive language. In: ICSCSS (2024)
21. Bowker, J., Ophoff, J.: Reducing exposure to hateful speech online. In: Science and Information Conference. Springer (2022)
22. Braun, T., Pekaric, I., Apruzzese, G.: Understanding the process of data labeling in cybersecurity. In: ACM/SIGAPP SAC. pp. 1596–1605 (2024)
23. Canossa, A., Salimov, D., Azadvar, A., Harteveld, C., Yannakakis, G.: For honor, for toxicity: Detecting toxic behavior through gameplay. In: CHI-PLAY (2021)
24. Chakrabarty, N.: A machine learning approach to comment toxicity classification. In: CIPR. Springer (2019)
25. Commission, E.: Sensitive data. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en
26. Cornel, J.A., Pablo, C.C., Marzan, J.A., Mercado, V.J., Fabito, B., Rodriguez, R., Octaviano, M., Oco, N., De La Cruz, A.: Cyberbullying detection for online games chat logs using deep learning. In: IEEE HNICEM (2019)
27. Corrêa, N.K.: Aira (2023). https://doi.org/10.5281/zenodo.6989727, https://github.com/Nkluge-correa/Aira

28. Costa-jussà, M., Meglioli, M., Andrews, P., Dale, D., Hansanti, P., Kalbassi, E., Mourachko, A., Ropers, C., Wood, C.: Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In: ACL (2024)
29. Deep, S., Singh, S., Yadav, S., Shedge, Y., Khade, A.: Creating an integrated browser plug-in for detecting and blocking obscene image/text content. In: International Conference on Networking and Communications (2024)
30. Dreier, L., Pirker, J.: Toxicity in twitch live stream chats: Towards understanding the impact of gender, size of community and game genre. In: IEEE CoG (2023)
31. d'Sa, A.G., Illina, I., Fohr, D.: Bert and fasttext embeddings for automatic detection of toxic speech. In: IEEE OCTA (2020)
32. Du Toit, J.L., Kotzé, E.: The automatic detection of abusive language in dota 2 chat messages. In: ACDSA (2024)
33. Ducasse, Q., Cotret, P., Lagadec, L., Stewart, R.: Benchmarking quantized neural networks on fpgas with finn. In: DATE Friday Workshop on System-level Design Methods for Deep Learning on Heterogeneous Architectures (2021)
34. ExplodingTopics: Amount of data created daily. Accessed: 6 Oct. 2025 (2025), https://explodingtopics.com/blog/data-generated-per-day
35. ExplodingTopics: How many gamers are there? Accessed: 6 Oct. 2025 (2025), https://explodingtopics.com/blog/number-of-gamers
36. ExplodingTopics: Internet traffic from mobile devices. Accessed: 6 Oct. 2025 (2025), https://explodingtopics.com/blog/mobile-internet-traffic
37. Falotico, R., Quatto, P.: Fleiss' kappa statistic without paradoxes. Quality & Quantity (2015)
38. Frommel, J., Sagl, V., Depping, A.E., Johanson, C., Miller, M.K., Mandryk, R.L.: Recognizing affiliation: Using behavioural traces to predict the quality of social interactions in online games. In: CHI (2020)
39. Ghosh, S., Lepcha, S., Sakshi, S., Shah, R.R., Umesh, S.: Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. In: Proc. Interspeech (2022)
40. Hanu, L., Unitary team: Detoxify. Github. https://github.com/unitaryai/detoxify
41. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E.: Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proc. ACL (2022)
42. Hsieh, C.Y., Li, C.L., YEH, C.K., Nakhost, H., Fujii, Y., Ratner, A.J., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Proc. ACL (2023)
43. Jahanbakhsh, F., Zhang, A.X., Karahalios, K., Karger, D.R.: Our browser extension lets readers change the headlines on news articles, and you won't believe what they did! Proceedings of the ACM on Human-Computer Interaction (2022)
44. Kazimierczak, M., Thanapattheerakul, T., Chan, J.H.: Enhancing security in whatsapp: a system for detecting malicious and inappropriate content. In: SOICT (2023)
45. Kilmer, E.D., Aslan, Z., Kowert, R.: Addressing toxicity and extremism in games: Conversations with the video game industry. Games and Culture (2024)
46. Kim, J., Wohn, D.Y., Cha, M.: Understanding and identifying the use of emotes in toxic chat on twitch. Online Social Networks and Media (2022)
47. Kordyaka, B., Karaosmanoglu, S., Laato, S.: Defining toxicity in multiplayer online games: A systematic literature review. Comp. Human Behavior Reports (2025)
48. Kou, Y.: Toxic behaviors in team-based competitive gaming: The case of league of legends. In: CHI-PLAY (2020)
49. Kwak, H., Blackburn, J., Han, S.: Exploring cyberbullying and other toxic behavior in team competition online games. In: CHI (2015)

50. Kwon, S., Liang, P., Tandon, S., Berman, J., Chang, P.j., Gilbert, E.: Tweety holmes: A browser extension for abusive twitter profile detection. In: CSCSC (2018)
51. Laato, S., Kordyaka, B., Hamari, J.: Traumatizing or just annoying? unveiling the spectrum of gamer toxicity in the StarCraft II community. In: CHI (2024)
52. Lobo, P.R., Daga, E., Alani, H.: Supporting online toxicity detection with knowledge graphs. In: AAAI ICWSM (2022)
53. Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., Semenov, N., Panchenko, A.: ParaDetox: Detoxification with parallel data. In: ACL (2022)
54. Mandryk, R.L., Frommel, J., Goyal, N., Freeman, G., Lampe, C., Vieweg, S., Wohn, D.Y.: Combating toxicity, harassment, and abuse in online social spaces: A workshop at chi 2023. In: Extended Abstracts of CHI (2023)
55. Märtens, M., Shen, S., Iosup, A., Kuipers, F.: Toxicity detection in multiplayer online games. In: NetGames (2015)
56. Merayo, N., Cotelo, R., Carratalá-Sáez, R., Andújar, F.J.: Applying machine learning to assess emotional reactions to video game content streamed on spanish twitch channels. Comp. Speech & Language (2024)
57. Miok, K., Nguyen-Doan, D., Škrlj, B., Zaharie, D., Robnik-Šikonja, M.: Prediction uncertainty estimation for hate speech classification. In: SLSP (2019)
58. Murnion, S., Buchanan, W.J., Smales, A., Russell, G.: Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security (2018)
59. Naseem, U., Shiwakoti, S., Shah, S.B., Thapa, S., Zhang, Q.: Gametox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In: ACL (2025)
60. Neto, J.A., Yokoyama, K.M., Becker, K.: Studying toxic behavior influence and player chat in an online video game. In: WI-IAT (2017)
61. Obadimu, A., Mead, E., Hussain, M.N., Agarwal, N.: Identifying toxicity within youtube video comment. In: SBP-BRiMS (2019)
62. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (2021)
63. Paul, H.L., Bowman, N.D., Banks, J.: The enjoyment of griefing in online games. Journal of Gaming & Virtual Worlds (2015)
64. Pawale, S., Pingat, V., Petkar, T., Patil, A., Waykar, S.: Toxiguard-a system that guides your vision towards toxic text. In: ICETI4T (2025)
65. Poeller, S., Dechant, M.J., Klarkowski, M., Mandryk, R.L.: Suspecting sarcasm: how league of legends players dismiss positive communication in toxic environments. In: CHI PLAY (2023)
66. Reid, E., Mandryk, R.L., Beres, N.A., Klarkowski, M., Frommel, J.: "bad vibrations": Sensing toxicity from in-game audio features. IEEE ToG (2022)
67. Schröer, S.L., Canevascini, N., Pekaric, I., Widmer, P., Laskov, P.: The dark side of the web: Towards understanding various data sources in cyber threat intelligence. In: 2025 EuroS&PW. pp. 79–89. IEEE (2025)
68. Sengün, S., Salminen, J., Jung, S.g., Mawhorter, P., Jansen, B.J.: Analyzing hate speech toward players from the mena in league of legends. In: ACM CHI (2019)
69. Shaer, O., Westendorf, L., Knouf, N.A., Pederson, C.: Understanding gaming perceptions and experiences in a women's college community. In: CHI (2017)
70. Shannaq, F., Hammo, B., Faris, H., Castillo-Valdivieso, P.A.: Offensive language detection in arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. IEEE Access (2022)

71. Sikiandani, N.M.D., Suarjaya, I.M.A.D., Putra, Y.P.: Browser-based detection of harmful content with deep learning model. J. Appl. Inf. Comp. (2025)
72. Sparrow, L.A., Galwey, R., Jovic, D., Hardwick, T., Butt, M.A.: Towards ethical ai moderation in multiplayer games. In: CHI PLAY
73. Statista: Esports - worldwide. Accessed October 6th 2025 (2024), https://www.statista.com/outlook/amo/esports/worldwide
74. Statista: Video game worldwide - statistics & facts. Accessed: 6 Oct. 2025 (2025), https://www.statista.com/topics/1680/gaming/
75. Steam: Community discussion, https://steamcommunity.com/discussions/
76. Stepanova, N., Muthemba, W., Todrzak, R., Cross, M., Ames, N., Raiti, J.: Natural language processing and sentiment analysis for verbal aggression detection; a solution for cyberbullying during live video gaming. In: PETRA (2021)
77. StreamHatchet: Most watched games on youtube gaming. Accessed: 6 Oct. 2025 (2025), https://streamhatchet.com/rankings/youtube/games/
78. Tang, W.Y., Reer, F., Quandt, T.: Investigating sexual harassment in online video games: How personality and context factors are related to toxic sexual behaviors against fellow players. Aggressive Behav. (2020)
79. Tricomi, P.P., Facciolo, L., Apruzzese, G., Conti, M.: Attribute inference attacks in online multiplayer video games: A case study on dota2. In: CODASPY (2023)
80. TwitchTracker: Twitch top streamers. Accessed: 6 Oct. 2025 (2025), https://twitchtracker.com/channels/ranking
81. Viggiato, M., Lin, D., Hindle, A., Bezemer, C.P.: What causes wrong sentiment classifications of game reviews? IEEE TOG (2021)
82. Vo, H.H.P., Tran, H.T., Luu, S.T.: Automatically detecting cyberbullying comments on online game forums. In: RIVF (2021)
83. Weld, H., Huang, G., Lee, J., Zhang, T., Wang, K., Guo, X., Long, S., Poon, J., Han, S.C.: Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection. arXiv preprint arXiv:2106.06213 (2021)
84. Wijesiriwardene, T., Inan, H., Kursuncu, U., Gaur, M., Shalin, V.L., Thirunarayan, K., Sheth, A., Arpinar, I.B.: Alone: A dataset for toxic behavior among adolescents on twitter. In: ICSI. Springer (2020)
85. Wijkstra, M., Rogers, K., Mandryk, R.L., Veltkamp, R.C., Frommel, J.: How to tame a toxic player? a systematic literature review on intervention systems for toxic behaviors in online video games. In: CHI-PLAY (2024)
86. Xue, S.S.: League of legends tribunal chatlogs (2020), https://www.kaggle.com/datasets/simshengxue/league-of-legends-tribunal-chatlogs
87. Yang, Z., Grenon-Godbout, N., Rabbany, R.: Game on, hate off: A study of toxicity in online multiplayer environments. ACM Games: Research and Practice (2024)
88. Zsila, Á., Shabahang, R., Aruguete, M.S., Orosz, G.: Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences. Aggressive Behavior (2022)

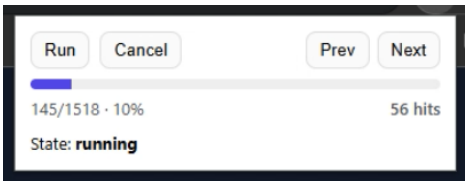

Fig. 3: The graphical interface of our browser extension. The analysis can be stopped or resumed.