thaulab@EEUCA 2026: Who Said What to Whom? A Targeting-Aware Neural-Symbolic Pipeline for Gaming Toxicity Detection

Anmol Guragain*, Marcos Estecha Garitagoitia,
Luis Fernando D'Haro Enríquez, Ricardo Córdoba
ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
*anmol.g@upm.es (Corresponding author)

Abstract

This paper describes our system for the EEUCA 2026 Shared Task on toxicity classification in gaming chat. We implement a three-stage pipeline combining an ensemble of two compact transformers (DeBERTa-v3-base, 184M; XLM-RoBERTa-base, 278M) with a Linguistically-Informed Mediator (LIM) that resolves inter-model disagreements through corpus-backed lexical normalization, class-conditional unigram scoring, multilingual profanity detection, and agentive targeting analysis grounded in speech act theory. The LIM specifically targets the minority classes (Hate & Harassment, Threats, and Extremism), which are the most safety-critical categories in real-world gaming moderation. To address the extreme class imbalance (1,450:1 Non-toxic to Extremism ratio), we introduce a two-stage data augmentation strategy using only the provided training data. Our system achieves a Macro F1 of 0.6441 and accuracy of 0.9062 on the official test set, ranking 3rd in Macro F1 and 1st in accuracy among all teams. The proposed pipeline is domain-portable: adapting to other gaming platforms requires substituting only the game-specific entity lexicon. Code is publicly available at https://github.com/Anmol2059/thaulab_EEUCA.

1 Introduction

Online gaming platforms host millions of real-time text interactions daily, and toxic behavior in these environments has been linked to serious consequences including cyberbullying, psychological harm, and player attrition (Parihar et al., 2021). A recent systematic review of 64 studies confirms that cyberbullying in multiplayer games is associated with anxiety, depression, and social withdrawal (Hu et al., 2025), and empirical evidence shows that toxic behavior propagates virally among teammates (exposure to toxic teammates increases a player's own toxicity likelihood by up to 30×), amplifying its reach when left undetected (Morrier et al., 2024).

The EEUCA 2026 Shared Task on Gaming Toxicity (Thapa et al., 2026; Hürriyetoğlu et al., 2026) introduces a six-class classification benchmark derived from World of Tanks chat logs (Naseem et al., 2025), annotated following the directed/undirected hate speech framework of Bhandari et al. (2023). The dataset poses three key challenges: extreme class imbalance (81.0% Non-toxic vs. 0.06% Extremism), multilingual content spanning 10+ languages, and domain-specific lexical ambiguity where violent vocabulary ("kill", "destroy") carries non-violent illocutionary force.

We implement a three-stage system combining neural ensemble classification with a rule-based Linguistically-Informed Mediator (LIM), following evidence that logical rules provide complementary signal to neural hate speech classifiers (Clarke et al., 2023; Awasthi et al., 2020). Our contributions are: (1) a two-stage augmentation strategy (confusion-pair-driven and contrastive boundary generation) that improves Macro F1 by +9.7% relative using only the provided data; (2) a LIM module grounded in speech act theory (Austin, 1962; Searle, 1969) that resolves ensemble disagreements through four interpretable, corpus-backed components; (3) empirical evidence that even multilingual transformers exhibit residual blind spots on domain-specific non-Latin profanity (22.6% of Hate & Harassment contains Cyrillic); and (4) demonstration that general-purpose toxicity models (toxic-bert) fail catastrophically in the gaming domain (Macro F1 = 0.3154), showing that gaming chat is a distinct linguistic register that requires domain-specific handling.

Related work. Recent NLP approaches to gaming toxicity include domain-adaptive pretraining of RoBERTa with match metadata for DOTA 2 and Call of Duty (Schurger-Foy et al., 2025), and hybrid

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 151–160

architectures combining LLM-generated embeddings with lightweight classifiers for Twitch moderation (Ansari et al., 2026). For class imbalance in hate speech detection, Zhang et al. (2024) show that focal loss (Lin et al., 2017) consistently yields peak performance, motivating our loss function choice. LLM-based data augmentation has proven effective for hate speech minority classes (Li et al., 2026), supporting our two-stage augmentation strategy (§2.1). The GameTox dataset (Naseem et al., 2025) additionally provides intent and slot filling annotations, but these labels were not released for the shared task, limiting participants to the six-class toxicity schema. Annotation disagreement is a recognized challenge in hate speech classification (Dehghan et al., 2025; Bhandari et al., 2023); we quantify its extent in this dataset in §D.

2 Task, Dataset, and Augmentation

The shared task (Thapa et al., 2026; Hürriyetoğlu et al., 2026) requires classifying World of Tanks chat into six categories: **Non-toxic** (0), **Insults** (1), **Other Offensive** (2), **Hate & Harassment** (3), **Threats** (4), and **Extremism** (5), evaluated by Macro F1. Table 1 shows the class distribution; the dataset contains 42,959 training, 5,367 validation, and 5,375 test samples with extreme imbalance (Non-toxic to Extremism ratio of 1,450:1). Messages are very short (median 2–4 tokens) and 6.9% contain Cyrillic script, rising to 22.6% in Hate & Harassment.

2.1 Two-Stage Augmentation

We augment minority classes without external data using a two-stage strategy (Figure 1; prompt templates in Appendix C).

Stage A uses a seed model (M0, DeBERTa-v3-base trained on original data) to identify confused class pairs: for each validation sample, we record the two highest-probability classes from M0 and flag the pair as a confusion boundary when the second-highest probability exceeds 0.15 (set empirically), indicating non-trivial model uncertainty between the two classes (Swayamdipta et al., 2020). This threshold was selected based on validation-set Macro F1 evaluated across candidate values $\{0.10, 0.15, 0.20, 0.25\}$; 0.15 maximized minority-class recall without introducing excess noise into the augmentation pool, as lower values produced near-duplicate confusion pairs while higher values missed meaningful boundary cases. Claude Opus 4.6 then generates synthetic samples targeting these confusion boundaries. This yields augmentation for Other Offensive, Hate & Harassment, and Threats.

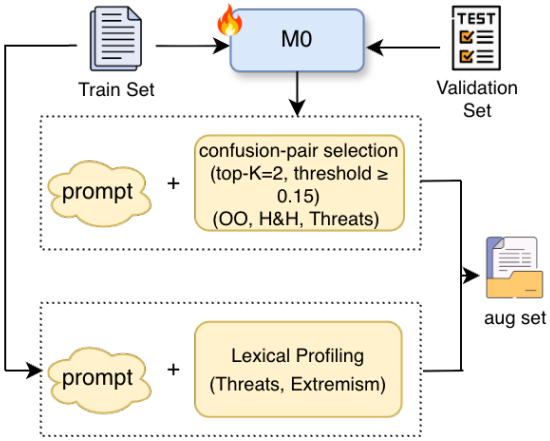

Figure 1: Two-stage augmentation. Stage A: confusion-pair-driven. Stage B: contrastive boundary with lexical profiling.

Class	Original	Added	Final
Non-toxic	34,797	0	34,797
Insults	5,925	0	5,925
Other Offensive	1,874	34	1,908
Hate & Harassment	279	155	434
Threats	60	235	295
Extremism	24	207	231
Total	42,959	631	43,590

Table 1: Training set class distribution before and after augmentation.

Stage B addresses Extremism (n=24) and supplemental Threats (n=60), which are too rare for confusion-pair analysis. We apply contrastive boundary augmentation: (1) mine class-discriminative tokens at $\geq 5\times$ frequency ratio using $P(c \mid w)$, the fraction of training messages containing word w that belong to class c (the same statistic used in the LIM's unigram scoring, §3.3.2); (2) generate cross-lingual variants and unmask leet-speak (e.g., `naz1`→`nazi`); (3) verify that generated samples fall within valid similarity bounds to real training data via cosine similarity in a TF-IDF subspace restricted to the mined discriminative vocabulary.

Table 1 shows the result: 631 synthetic samples, improving Macro F1 by +9.7% relative over M0.

3 System Architecture

Our system is a three-stage pipeline (Figure 2); hyperparameters for all models are in Appendix A.

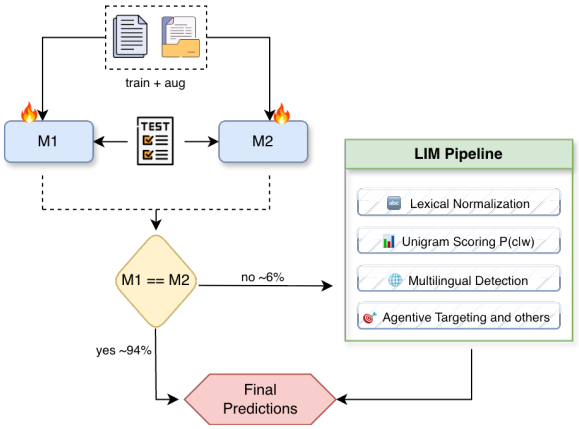

Figure 2: System pipeline. M1 (DeBERTa) and M2 (XLM-R) produce predictions; agreements (∼94%) are accepted, disagreements (∼6%) are refined by the LIM.

Model	F1	Acc	Prec	Rec
M0: DeBERTa (seed)	.5234	.8945	.5463	.5156
M1: DeBERTa (aug.)	.5742	.8979	.6042	.5538
M2: XLM-RoBERTa	.5613	.8910	.5670	.5730
M3: BERT-base	.5439	.8947	.5252	.5811
M4: toxic-bert	.3154	.7518	.3187	.5497

Table 2: Individual model results. M0/M1 share the DeBERTa architecture (seed vs. augmented). M1 and M2 form the final ensemble.

3.1 Stage 1: Model Exploration and Selection

We explored four base-sized transformers, all trained with focal loss (Lin et al., 2017) (γ=2.0, α=None). Table 2 summarizes results. We select M1 (DeBERTa-v3-base (He et al., 2023), 184M, highest F1 after augmented training) and M2 (XLM-RoBERTa (Conneau et al., 2020), 278M, complementary multilingual coverage) for the ensemble. M3 (BERT-base (Devlin et al., 2019), 110M) served as a development baseline. M4 (toxic-bert (Hanu and Unitary team, 2020), 110M, frozen backbone + MLP) achieved only 0.3154 Macro F1 despite toxicity-specific pre-training, demonstrating that general-domain toxicity representations do not transfer to gaming contexts, where violent vocabulary is routinely non-toxic and multilingual slang is pervasive. This is part of what the LIM's domain-specific linguistic rules are designed to address.

3.2 Stage 2: Agreement-Based Fusion

When M1 and M2 agree (∼94%), we accept the consensus. For the ∼6% disagreements, we adopt the prediction with higher softmax probability as the initial estimate and route to the LIM.

3.3 Stage 3: Linguistically-Informed Mediator

The LIM refines disagreement predictions through four sequential components. It combines neural and symbolic processing: the ensemble captures distributional semantics, while the LIM encodes domain-specific linguistic facts that neural models cannot reliably learn from limited minority-class data. Every LIM decision traces back to a specific rule and corpus statistic, making it auditable.

3.3.1 Corpus-Backed Lexical Normalization

We normalize test messages (lowercase, strip punctuation, collapse expressive lengthening (Brody and Diakopoulos, 2011): "hahaha"→"haha") and perform exact-match lookup against train+val. Matches with $\geq$ 2 occurrences and $\geq$ 60% majority agreement adopt the majority label. These conservative thresholds directly reflect the annotation noise quantified in §D.

3.3.2 Class-Conditional Unigram Scoring

Inspired by the token-level analysis of Naseem et al. (2025), we compute $P(c \mid w) = \frac{n(w,c)}{n(w)}$ for each word w in the training vocabulary, where c is a class label, $n(w, c)$ is the number of messages containing w in class c, and $n(w)$ is the total count, effectively a unigram Naïve Bayes estimate. Words exceeding a precision threshold $P(c \mid w) \geq 0.80$ with sufficient support ($n(w) \geq 5$) serve as high-confidence minority-class indicators (Wiegand et al., 2018). For instance, identity-based slurs consistently map to H&H (P=1.00), while "kys" maps to Threats and leet-speak variants like "naz1" to Extremism. Overrides apply only toward safety-critical classes (3–5: H&H, Threats, Extremism), prioritizing precision to avoid false escalation.

3.3.3 Multilingual Profanity Detection

While M2 (XLM-RoBERTa) handles multilingual tokenization, our validation analysis revealed that *both* M1 and M2 still misclassify domain-specific non-Latin profanity, particularly terms rare even in XLM-R's 100-language pre-training. We applied the same statistics to non-Latin tokens, flagging words where $P(\text{toxic} \mid w) = 1 - P(\text{Non-toxic} \mid w) \geq 0.80$ but both models predicted Non-toxic. This yielded an empirically-validated multilingual lexicon organized by language family: East Slavic (Russian, Ukrainian), West Slavic (Polish, Czech), and other (Turkish, Hungarian, German). Reclas-

sification follows targeting: player-directed → Insults; game-directed → Other Offensive.

3.3.4 Agentive Targeting and Pragmatic Refinement

Drawing on speech act theory (Austin, 1962; Searle, 1969), we formalize the targeting function $\tau(m)$. Let T denote the set of tokens flagged as toxic by the preceding LIM components (unigram scoring and multilingual detection). For a message m containing a toxic token $t \in T$:

$$\tau(m) = \begin{cases} \textsc{Other-dir} & \text{if } \exists\, p \in P_2 : p \prec t \\ \textsc{Self-dir} & \text{if } \exists\, p \in P_1 : p \prec t \\ \textsc{Entity-dir} & \text{if } \exists\, e \in E : e \prec t \\ \textsc{Untargeted} & \text{otherwise} \end{cases} \tag{1}$$

where P_2/P_1 are second/first-person pronoun sets (English and Russian), E is a Game-Specific Entity (GSE) lexicon covering vehicles (400+ tanks), mechanics (*rng*, *arty*, *cap*), map locations (*Himmelsdorf*, *hill*, *banana*), and game roles (*light*, *heavy*, *TD*), and $\prec$ denotes linear precedence in the message. Table 3 maps targeting types to labels.

$\tau(m)$	Signal	$\hat{y}$
Other-dir	$P_2 + T$ (*you* + insult)	Insults
Self-dir	$P_1 + T$ (*I* + insult)	Non-toxic
Entity-dir	$E + T$ (*arty* + profanity)	Other Off.
Untargeted	T only	Non-toxic

Table 3: Targeting function $\tau(m)$. E = GSE lexicon. Based on speech act theory (Searle, 1969).

This component also applies censored-text recovery ([GSE] + [***] → Non-toxic) and implicit word sense disambiguation: "kill that Tiger" (GSE → Non-toxic) vs. "kill yourself" (person → Threats) (Firth, 1957).

4 Results and Discussion

Table 4 shows incremental results on the official test set. The ensemble improves over the best single model through complementary coverage, and the LIM further refines the ∼6% disagreements, with the largest contribution from unigram scoring (Table 4). The LIM's impact is concentrated in safety-critical minority classes, where high-precision corrections ensure that identity-based hate, threats, and extremist content are not missed by the neural ensemble.

Our system ranks 3rd in Macro F1 (0.6441) but achieves the **highest accuracy** (0.9062) among all participating teams, indicating the fewest total errors; the F1 gap to the top-ranked systems (0.7041, 0.6725) is concentrated in minority class recall. **Annotation noise** accounts for part of this ceiling: 340 unique messages carry conflicting labels across 7,416 training samples (17.3%), with the Non-toxic ↔ Insults boundary alone responsible for 6,455 conflicting samples, reflecting the same ambiguity between a playful insult and a genuine attack that makes this boundary the hardest to learn. **Multilingual blind spots** persist even with XLM-RoBERTa: domain-specific Cyrillic profanity is concentrated at 22.6% of H&H messages (3.5× the dataset average), reflecting the prevalence of Russian and Ukrainian identity-based slurs that fall outside standard multilingual pre-training corpora. **Domain gap**: the toxic-bert result (F1 = 0.3154 vs. 0.5439 for vanilla BERT) shows that Twitter/Reddit toxicity pre-training actively hurts gaming performance by associating GSE terms with toxicity.

System	F1	Acc	Prec	Rec
Best single (M1)	.5742	.8979	.6042	.5538
M1+M2 ensemble	.6032	.9059	.5713	.6579
+ Lex. norm.	.6107	.9057	.5782	.6591
+ Unigram scoring	.6221	.9044	.5964	.6559
+ Multilingual	.6256	.9047	.5970	.6626
+ Targeting & ref.	**.6441**	**.9062**	**.6334**	**.6601**

Table 4: Incremental ablation. Top: models. Bottom: LIM components applied to ∼6% disagreements. Full LIM includes targeting, censored-text recovery, and boundary enforcement.

We presented **thaulab**'s system for the EEUCA 2026 GameTox Shared Task, achieving Macro F1 of 0.6441 (3rd) and the highest accuracy (0.9062) using exclusively base-sized models and no external data. The pipeline is adaptive by design: augmentation targets the boundaries the model struggles with, and the LIM concentrates corrections on Extremism, Threats, and H&H, categories where misclassification causes real harm beyond any leaderboard metric. The three-stage framework generalizes to other gaming platforms with only GSE lexicon substitution. Key findings: (1) symbolic mediation on ensemble disagreements improves safety-critical minority-class detection; (2) multilingual transformers retain blind spots on domain-specific profanity; (3) agentive targeting distinguishes toxic intent from benign game communication; (4) general toxicity models fail in gaming contexts.

Limitations

The LIM relies on simple symbolic methods (lexicon lookup, unigram statistics, pronoun-based targeting) rather than learned components. While this trades off flexibility for explainability, since every correction traces back to a specific corpus statistic or linguistic pattern, it likely leaves performance on the table. We set strict thresholds throughout (e.g., $P(c \mid w) \geq 0.80$, majority agreement $\geq 60\%$) and have not yet explored relaxing them per class or per confidence region. The four LIM components are applied in a fixed sequential order without searching over alternative orderings, and the LIM currently operates only on the $\sim$6% of samples where M1 and M2 disagree; extending symbolic mediation to the full prediction set could improve coverage but risks false positives at the noisy majority-class boundaries. On the augmentation side, all synthetic data was generated using Claude Opus 4.6, a safety-aligned model that may undergenerate realistic toxic content due to its content filtering; less restricted models (e.g., Grok) could potentially yield more naturalistic minority-class samples, particularly for Extremism where only 24 original samples exist. Finally, the multilingual lexicon is manually curated and incomplete for unrepresented languages, the GSE lexicon is specific to World of Tanks, targeting analysis cannot resolve implicit targeting or zero pronouns in pro-drop languages, and we could not experiment with larger backbone models due to time and compute constraints.

Ethics Statement

This work classifies toxic language including slurs and profanity in multiple languages. Examples are reported solely for scientific reproducibility. Our multilingual profanity lexicon is compiled for research purposes only.

Acknowledgements

This work is supported by the project BRAINS (PID2024-155948OB-C52), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe" and by the "European Union"; and by the Program *Actividades I+D Procesos Humanos y Sociales* of the Comunidad de Madrid, Spain, under Project PHS2024/PH-HUM-52 (Innovatrad-CM).

References

Baktash Ansari, Elias Martin, and Afra Mashhadi. 2026. ToxiTwitch: Toward emote-aware hybrid moderation for live streaming platforms. *Preprint*, arXiv:2601.15605.

John L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. 2020. Learning from rules generalizing labeled exemplars. In *International Conference on Learning Representations*.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. CrisisHateMM: Multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Somaiyeh Dehghan, Mehmet Umut Sen, and Berrin Yanikoglu. 2025. Dealing with annotator disagreement in hate speech classification. *Preprint*, arXiv:2502.08266.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*, pages 4171–4186.

John R. Firth. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*, pages 1–32.

Laura Hanu and Unitary team. 2020. Detoxify. https://huggingface.co/unitary/toxic-bert.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Yunhao Hu, Sophie Evelyn, and Elizabeth M. Clancy. 2025. Player versus player: A systematic review of cyberbullying in multiplayer online games. *Computers in Human Behavior*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, Laxmi Thapa, and Surabhi Adhikari. 2026. Overview of the workshop on event extraction and understanding: Challenges and applications. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Peiran Li, Jan Fillies, and Adrian Paschke. 2026. ToxiGAN: Toxic data augmentation via LLM-guided directional adversarial generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Jacob Morrier, Amine Mahmassani, and R. Michael Alvarez. 2024. Uncovering the viral nature of toxicity in competitive online video games. *Preprint*, arXiv:2410.00978.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. Gametox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Adrien Schurger-Foy, Rafal Dariusz Kocielnik, Caglar Gulcehre, and R. Michael Alvarez. 2025. Context-aware toxicity detection in multiplayer games: Integrating domain-adaptive pretraining and match metadata. *Preprint*, arXiv:2504.01534.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026. Understanding toxic behavior in gaming communities using AI to promote healthier digital spaces. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Yaqi Zhang, Viktor Hangya, and Alexander Fraser. 2024. A study of the class imbalance problem in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*. Association for Computational Linguistics.

A Hyperparameter Configuration

Table 5 lists the model-specific training configuration for all four architectures. The pre-trained checkpoints are `microsoft/deberta-v3-base` (M0/M1), `xlm-roberta-base` (M2), `bert-base-uncased` (M3), and `unitary/toxic-bert` (M4). All models share the same base settings: focal loss with focusing parameter $\gamma = 2.0$ and no class balancing ($\alpha =$ None), maximum sequence length of 64 tokens, batch size 32, 5 training epochs, AdamW optimizer with weight decay 0.01, and random seed 42.

While M2 and M3 natively instantiate in single-precision (`float32`), the DeBERTa-v3 checkpoints (M0/M1) are natively stored in half-precision (`float16`). We observed that fine-tuning DeBERTa-v3 in `float16` resulted in catastrophic gradient collapse (`NaN` loss) during the initial training steps, a known instability caused by arithmetic overflow within DeBERTa's Disentangled Attention matrices, where intermediate activation values exceed the `float16` maximum representable limit. To resolve this, we explicitly upcast the DeBERTa weights to `float32` during initialization, providing sufficient numerical stability for the attention mechanism to converge. We additionally tested $\gamma = 2.5$, dynamic α (inverse class frequency), and a two-stage hierarchical approach (binary toxic/non-toxic classification followed by fine-grained 6-class prediction within the toxic branch). All alternatives yielded marginal differences (Δ Macro F1 < 0.005), so we standardized the simplest configuration for reproducibility across all architectures.

B LIM Component Details

Table 6 lists all LIM thresholds, selected on the validation set and held fixed during test evaluation. The lexical normalization majority threshold was set at 60% rather than 50% because lower values introduced false corrections at the noisy Insults $\leftrightarrow$ Other Offensive boundary. The unigram precision cutoff of $P \geq 0.80$ was chosen because at $P \geq 0.70$, ambiguous terms (e.g., "monkey" at $P(\text{H\&H})=0.75$) triggered false positives; raising to 0.80 retains only unambiguous high-precision tokens. These thresholds are intentionally strict for the competition setting and can be relaxed for higher-recall deployment.

Throughout the LIM, we enforce annotation-guideline boundaries (Naseem et al., 2025): identity-based slurs $\rightarrow$ H&H; 2nd person + non-identity insult $\rightarrow$ Insults; profanity without personal targeting $\rightarrow$ Other Offensive; game callouts and GSE terms $\rightarrow$ Non-toxic; directed violence + personal target $\rightarrow$ Threats; political ideology and recruitment $\rightarrow$ Extremism.

C Augmentation Pipeline

Toxic vocabulary mining. Class-conditional unigram probabilities $P(c \mid w)$ are computed as described in §3.3.2. For augmentation, we additionally flag *class-discriminative tokens* using the frequency ratio $\frac{n(w,c)/N_c}{n(w)/N} \geq 5$, where N_c and N are class and corpus sizes respectively; this yields a focused toxic vocabulary substantially smaller than the full $\sim$30K vocabulary, defining the TF-IDF subspace for similarity gating below.

Before computing any statistics, all text undergoes leet-speak normalization to unmask common obfuscation patterns prevalent in gaming chat. The character substitution mappings are: `0`$\rightarrow$`o`, `1`$\rightarrow$`i`, `3`$\rightarrow$`e`, `4`$\rightarrow$`a`, `5`$\rightarrow$`s`, `7`$\rightarrow$`t`, `@`$\rightarrow$`a`, `$`$\rightarrow$`s`. This normalization is applied consistently in both the augmentation pipeline (for seed term selection and similarity verification) and the LIM (for unigram scoring and multilingual detection at inference time).

Cosine similarity gating in the toxic subspace. To verify that generated samples are linguistically consistent with real training data, we project both real and synthetic utterances into a *toxic-only TF-IDF subspace*. Rather than computing TF-IDF vectors over the full $\sim$30K vocabulary (which produces extremely sparse, high-dimensional vectors for short gaming messages of 2–4 tokens), we restrict the vocabulary to only the mined class-discriminative terms. This projection substantially reduces dimensionality and eliminates the sparsity problem inherent in full-vocabulary TF-IDF for short texts. Cosine similarity between each generated sample and its nearest real training neighbor in this subspace serves as a geometric filter: samples that fall below a minimum similarity threshold are rejected as out-of-distribution, while samples above a maximum threshold are rejected as near-duplicates of existing training data. This dual-threshold approach ensures that generated samples are close enough to the training distribution to be realistic, yet sufficiently novel to provide genuine augmentation value.

Parameter	M0/M1 (DeBERTa-v3)	M2 (XLM-RoBERTa)	M3 (BERT)	M4 (toxic-bert)
Total Parameters	184M	278M	110M	110M
Trainable Parameters	184M (full)	278M (full)	110M (full)	∼200K (MLP only)
Training Paradigm	Full fine-tune	Full fine-tune	Full fine-tune	Frozen backbone + MLP
Learning Rate	1×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-3} (MLP)
Warmup Steps	10% of total	0	0	0
Weight Initialization	`float32` (upcast)	`float32` (native)	`float32` (native)	`float32` (native)

Table 5: Model-specific training hyperparameters. M0 and M1 share the identical configuration; M0 is trained on the original training set, M1 on the augmented set.

Component	Parameter	Value
Lex. Norm.	Min occurrences	≥ 2
	Majority threshold	$\geq 60\%$
	Normalization	Lower, strip, dedup (≥ 3)
Unigram	$P(c \mid w)$ cutoff	≥ 0.80
	Min support $n(w)$	≥ 5
	Direction	Minority ↑ only
	Freq. ratio	$\geq 5\times$ baseline
Multilingual	Languages	10+
	Mining criterion	$P(\text{toxic} \mid w) \geq 0.80$ + both models wrong
	Reclassification	Targeting-sensitive
Targeting	Pronoun sets	EN, RU
	GSE lexicon E	Vehicles (400+), mechanics, maps, roles
	Censored pattern	[GSE] + [***]

Table 6: LIM thresholds. All selected on the validation set.

Generation API configuration. All synthetic samples were generated using the Claude Opus 4.6 API (`claude-opus-4-6-20250514`). We used a temperature of `1.0` to encourage lexical diversity across generated samples, `max_tokens = 2048`, and `top_p = 1.0` (no nucleus truncation). No additional system-level parameters were set beyond the defaults; the full generation behavior is governed solely by the prompt templates below. These settings are fixed across both Template A and Template B calls to ensure reproducibility.

Template A: Confusion-pair-driven generation. For classes identified through the seed model's (M0) prediction uncertainty, we provide the language model with the target class definition, representative seed examples from the training set, and the specific confused class pair that the model struggles with:

```
You are a data augmentation assistant
for a toxicity classification dataset
derived from World of Tanks in-game
chat. Your task is to generate realistic
synthetic chat messages for a specific
toxicity class.

Target class: {CLASS_NAME}
Class definition: {CLASS_DEFINITION}

The class definitions follow the
annotation guidelines from the GameTox
dataset (Naseem et al., 2025):
- Hate and Harassment: Identity-based
hate or harassment (racism, sexism,
homophobia)
- Threats: Threats of violence, physical
safety, terrorism, or doxxing
- Extremism: Extremist views,
grooming/recruitment for extremist
groups
- Insults and Flaming: Insults or
attacks not based on identity
- Other Offensive: Offensive content
not covered by the above categories
- Non-toxic: Neutral game communication

Seed examples from the training data:
{SEED_EXAMPLES}

Confused with: {CONFUSED_CLASS}
(our classifier frequently confuses
{CLASS_NAME} with {CONFUSED_CLASS})

Requirements:
1. Generate exactly 20 new chat messages
that CLEARLY belong to {CLASS_NAME} and
NOT to {CONFUSED_CLASS}.
2. Each message should be 1-8 words long
(typical length in game chat).
3. Include common gaming abbreviations,
slang, and informal spelling.
4. Include multilingual variants where
appropriate (Russian, Polish, Turkish,
German).
5. Each message must be unambiguously
classifiable by a human annotator
following the guidelines above.
6. Do NOT repeat or closely paraphrase
the seed examples.
7. Output one message per line with no
numbering or formatting.
```

Template B: Contrastive boundary augmentation. For extreme minority classes (Extremism with only $n = 24$ training samples, and supplemental Threats with $n = 60$) that are too rare to appear reliably in confusion-pair analysis, we provide discriminative keywords mined from the training set along with explicit instructions to generate boundary-proximal samples:

```
You are a data augmentation assistant for
```

```
a toxicity classification dataset from
World of Tanks in-game chat.

Target class: {CLASS_NAME}
Class definition: {CLASS_DEFINITION}
Adjacent (easily confused) class:
{ADJACENT_CLASS}
Adjacent class definition:
{ADJACENT_DEFINITION}

Discriminative keywords for {CLASS_NAME}
(statistically mined from training data,
P(class|word) >= 0.80):
{HIGH_P_KEYWORDS}

Existing training examples of
{CLASS_NAME}:
{SEED_EXAMPLES}

Requirements:
1. Generate exactly 20 new messages
that belong to {CLASS_NAME}.
2. CRITICAL: Messages must be
CLOSE to the decision boundary with
{ADJACENT_CLASS}. They should be
challenging to classify, but still
clearly {CLASS_NAME} according to the
annotation guidelines.
3. Include cross-lingual variants
(Russian, Polish, Turkish, German).
4. Vary message length (1-8 words).
5. Each message should be
distinguishable from {ADJACENT_CLASS}
ONLY by the specific class-defining
linguistic feature (e.g., identity
targeting for H&H vs. skill targeting
for Insults, or political ideology for
Extremism vs. identity hate for H&H).
6. Do NOT repeat seed examples.
7. Output one message per line with no
numbering.
```

Generation results and quality control. Template A yielded 34 Other Offensive, 155 Hate & Harassment, and a portion of the Threats samples. Template B yielded all 207 Extremism samples and supplemental Threats samples, for a combined total of 631 synthetic samples. All generated samples underwent three quality control steps: (1) cosine similarity gating in the toxic TF-IDF subspace to reject out-of-distribution and near-duplicate generations; (2) exact and near-duplicate removal against the original training set to prevent data leakage; (3) manual spot-checking of a random 10% subset for label consistency with the annotation guidelines. Class definitions in both templates were drawn directly from the annotation guidelines of Naseem et al. (2025).

D Annotation Noise Analysis

A key challenge in the GameTox dataset is annotation inconsistency at class boundaries. We identify 340 unique normalized messages that appear with conflicting labels across their multiple occurrences in the training set, collectively affecting 7,416 individual training samples (17.3% of the dataset). This inconsistency arises because identical text appears in different game sessions and receives different annotations each time. For instance, a player typing "wtf" in one match may be reacting to an unfair death (Other Offensive), while in another match the same message is interpreted as a neutral exclamation (Non-toxic). This is not a failure of multiple annotators disagreeing on a single instance; rather, it reflects the genuine context-dependence of short gaming messages.

Table 7 shows representative examples. The column n indicates the total number of times that normalized message appears in the training set across all game sessions. The label distribution shows the percentage of those n occurrences assigned to each class.

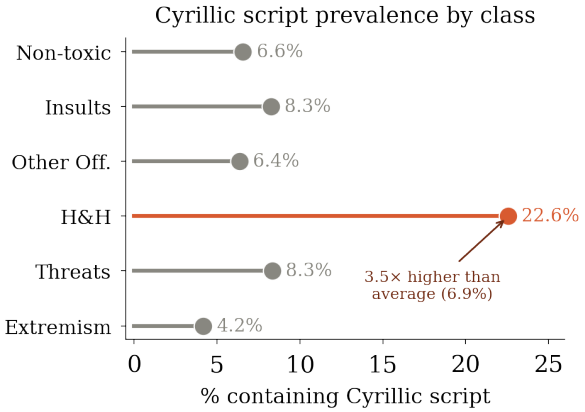

Figure 3: Cyrillic-script prevalence by toxicity class. Hate & Harassment contains 3.5× more Cyrillic content than the dataset average (6.9%), indicating that non-Latin-script profanity is structurally concentrated in the most severe toxicity category.

Message	n	Label Distribution
gg	2,703	NT: 99.9%, Ins: 0.1%
wtf	208	OO: 77.4%, NT: 22.1%, Ins: 0.5%
cap	192	NT: 96.4%, OO: 2.1%, Ins: 1.6%
arty	189	NT: 97.4%, Ins: 2.6%
idiot	101	Ins: 94.1%, NT: 5.0%, OO: 1.0%
ffs	55	OO: 80.0%, NT: 16.4%, Ins: 3.6%

Table 7: Annotation noise examples. n = total occurrences in training data. The same normalized text receives different labels across game sessions. NT = Non-toxic, Ins = Insults, OO = Other Offensive.

Figure 4 visualizes the magnitude of annotation conflicts across all class pairs. Each bubble represents a pair of classes; the bubble size and color intensity are proportional to the num-

ber of training samples where the same message receives labels from both classes. The Non-toxic ↔ Insults boundary dominates with 6,455 conflicting samples, reflecting the fundamental ambiguity between a playful insult and a genuine attack in gaming chat. The Non-toxic ↔ Other Offensive boundary (1,528 samples) and the Insults ↔ Other Offensive boundary (958 samples) are the next most noisy. Notably, minority class boundaries (involving H&H, Threats, or Extremism) exhibit minimal noise, because the linguistic signals for these classes (identity-based slurs, directed violence, political ideology) are more distinctive and less context-dependent.

This noise pattern directly informs the LIM design: we use conservative thresholds ($\geq 60\%$ majority agreement) for the noisy majority-class boundaries, while applying more aggressive corrections for minority classes where annotation agreement is near-unanimous.

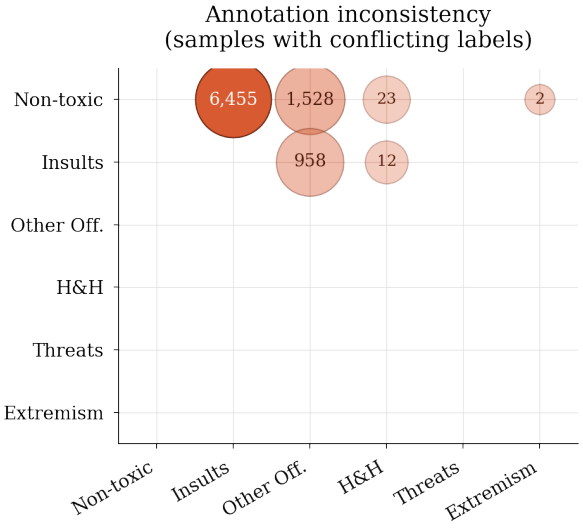

Figure 4: Annotation inconsistency across all class pairs. Bubble size reflects the total number of training samples where identical normalized messages receive conflicting labels from the two classes. The Non-toxic ↔ Insults boundary dominates at 6,455 conflicting samples, illustrating the context-dependent nature of short gaming messages.

E Multilingual Content Analysis

Figure 3 shows the proportion of Cyrillic-script messages by toxicity class. The 22.6% concentration in Hate & Harassment (compared to a 6.9% dataset average, a 3.5× difference) reflects the prevalence of Russian and Ukrainian identity-based profanity systems, collectively known as *mat*, which include some of the strongest and most targeted slurs in the Slavic language family. Beyond Cyrillic, the dataset contains content in Polish and Czech (0.30% of training data), Turkish (0.31%), and Hungarian (0.30%). In the test set, 389 messages (7.2%) contain Cyrillic script, 22 contain Turkish characters, and 17 contain Polish/Czech characters.

Why XLM-RoBERTa is insufficient. A natural question is why the LIM's multilingual lexicon is needed given that M2 (XLM-RoBERTa) is pre-trained on 100 languages including Russian, Ukrainian, Polish, and Turkish. The answer lies in the distinction between *general-vocabulary* multilingual competence and *domain-specific* profanity detection. XLM-RoBERTa's pre-training corpus (CommonCrawl) contains formal and semi-formal text, but underrepresents the specific register of gaming chat profanity: context-dependent slurs that are used as identity-based attacks in one context and as general frustration in another, obfuscated forms of profanity, and compound insults that combine multiple languages within a single utterance. Our validation analysis confirmed this empirically: we identified tokens where $P(\text{toxic} \mid w) = 1 - P(\text{Non-toxic} \mid w) \geq 0.80$ in the training data, yet *both* M1 (DeBERTa) and M2 (XLM-RoBERTa) predicted Non-toxic on the validation set. The LIM's multilingual lexicon targets precisely these residual blind spots, not as a replacement for XLM-RoBERTa's multilingual capacity, but as a domain-specific complement to it.