

GChat-Detect: Domain-Adaptive Toxicity Detection for Online Gaming Chats Using DistilBERT

Syeda Seerat Zahra Naqvi
ORCID: 0009-0003-8819-2464
Dept. of Data Science University of Kotli, AJ&k

Hameedur Rahman
ORCID: 0000-0001-8892-9911
Dept. of Computer Games Development Air University, Islamabad

Junaid Tariq
ORCID: 0000-0001-9619-2649
Dept. of Computer Science NUML, Islamabad

Corresponding Author: Syeda Seerat Zahra Naqvi*

Received: 08-02-2026

Revised: 01-03-2026

Accepted: 09-03-2026

Published: 24-03-2026

ABSTRACT

Online multiplayer gaming platforms host millions of concurrent users, making them fertile ground for toxic communication, hate speech, and harassment. Existing content moderation systems struggle with gaming-specific slang, abbreviations, and context-dependent expressions that are absent from general-purpose natural language processing (NLP) corpora. This study, proposes GChat-Detect, a fine-tuned DistilBERT model enhanced with a game-domain adaptive vocabulary layer trained on the publicly available Jigsaw Toxic Comment Classification Challenge dataset. This paper introduces a GameAdapt pre-processing pipeline that handles gaming abbreviations, emote sequences, and code-mixed chat language. This work deploys the model that achieves 93.4% accuracy, 92.8% macro F1-score, and an AUC-ROC of 0.974, outperforming baseline approaches including Logistic Regression, SVM, BiLSTM, and BERT-base. An ablation study confirms that domain-adaptive fine-tuning contributes a 2.2% improvement in F1-score over generic fine-tuning. The proposed framework offers a real-time deployable solution for toxicity detection in online gaming environments relevant to emerging digital economies.

Keywords: Toxicity Detection, Online Gaming, DistilBERT, Natural Language Processing, Hate Speech, Transfer Learning, Game Domain Adaptation

INTRODUCTION

The online gaming sector in the world grew to more than 3.2 billion active gamers and earned a revenue of more than 200 billion in 2024 alone [1]. PUBG Mobile, League of Legends, Fortnite and Call of Duty are examples of multiplayer online games that produce billions of in-game chat messages every day. On the one hand, these platforms bring about social interaction, community building, but on the other hand, these platforms are used to spread toxic behavior such as harassment, hate speech, racial slurs, threats, and cyberbullying. In a 2024 report by the Anti-Defamation League, it was found that 74 percent of online gamers were subjected to some type of harassment within the last year, with 53 percent reporting more serious toxicity like identity-based hate speech has been presented to them in the last year. Gaming content moderation poses its own set of challenges that make it different to the overall social media toxicity detection. First, gaming chats are dominated by domain-specific language abbreviations such as 'gg' (good game), 'noob' (inexperienced player), 'rekt' (defeated), and 'inting' (intentionally feeding) — which standard NLP models fail to interpret correctly. Second, rapid-fire, short-form messages with deliberate misspellings and character substitutions (e.g., 'st*pid', 'b@d') are commonly used to evade rule-based filters. Third, the multilingual and code-mixed nature of global

gaming communities adds another layer of linguistic complexity [2]. Existing automated approaches to toxicity detection predominantly rely on rule-based keyword filtering [3], classical machine learning classifiers [4], or general-domain transformer models that are not adapted to the gaming context [5]. These systems suffer from high false-positive rates when flagging gaming jargon as toxic and high false-negative rates when encountering obfuscated hate speech. In this paper, this study proposes GChat-Detect, a gaming-domain adapted toxicity detection framework built on Dis-tilBERT, a lightweight, efficient variant of BERT that retains 97% of BERT’s language understanding capability while being 40% smaller and 60% faster [6]. Our key contributions are as follows:

- This study proposes a GameAdapt preprocessing pipeline specifically designed for online gaming chat language, including gaming abbreviation normalization, emote han-dling, and obfuscation detection.
- This work fine-tunes DistilBERT on the Jigsaw Toxic Comment Classification Challenge dataset (Kaggle, 165,198 samples) with gaming-domain vocabulary aug-mentation.
- This study conducts a comprehensive ablation study val-idating the contribution of each pipeline component.
- Our model achieves state-of-the-art performance among computationally efficient models, making it suitable for real-time deployment on gaming servers.

RELATED WORK

Traditional Approaches to Toxicity Detection

Early work on online toxicity detection relied heavily on lexicon-based methods and rule-based systems. Nobata et al. developed a comment scoring system using linguistic features combined with n-gram language models, achieving competi-tive results on Yahoo news data [3]. However, such approaches fail systematically when confronted with adversarial obfusca-tion. Schmidt and Wiegand provided a comprehensive survey of hate speech detection, noting that SVM classifiers with TF-IDF features dominated the field prior to the deep learning era, with accuracy plateauing around 82–85% on benchmark datasets [4]

TABLE I: Comparative Analysis of Hate Speech Detection Models and Identified Gaps

Ref.	Title	Model	Dataset	Results	Limitations
[7]	Survey of Transformer-Based Mod-els for Toxic Language Detection (2024)	Survey: BERT, RoBERTa variants	Multiple	–	No gaming-domain coverage; general NLP focus only
[8]	Context-Aware Hate Speech Detec-tion using Fine-Tuned BERT and Rule-Based Features (2024)	Fine-tuned BERT + keyword rules	Twitter, Reddit	88.4	No leet-speak or obfuscation handling; lacks gaming context
[9]	Multilingual Hate Speech Detection using mBERT for Low-Resource Languages (2024)	mBERT multilingual	Urdu social media	84.1	Urdu-only focus; limited transferability to gaming chat
[10]	Hate Speech Detection using BERT Fine-Tuning on Large-Scale	BERT-base fine-tuned	Jigsaw (Kaggle)	89.7	No domain adaptation; lacks gaming vocabulary;

	Toxicity Datasets (2023)				computationally heavy
[11]	Attention-Based BiLSTM for Offensive Language Detection with GloVe Embeddings (2023)	BiLSTM + attention + GloVe	HatEval, OffComEval	85.2	Static embeddings; ineffective for gaming slang and character substitution
[12]	Behavioral Toxicity Detection in Online Multiplayer Games (2023)	Behavioral feature analysis	League of Legends logs	79.0	Relies on behavioral signals; lacks NLP-based chat analysis
[13]	Lightweight Toxicity Detection using DistilBERT (2023)	DistilBERT (general)	Wikipedia, Twitter	90.3	No domain ad gaming specific vocabulary
[14]	Cross-Lingual Offensive Language Detection using XLM-RoBERTa (2023)	XLM-RoBERTa cross-lingual	SemEval multilingual	86.8	Cross-lingual focus; not evaluated in gaming environments
[15]	Large Language Models for Offensive Content Detection: GPT-4 Benchmarking (2024)	GPT-4 zero-shot + fine-tune	OffComEval 2024	91.2	Extremely high compute; not suitable for real-time deployment

Deep Learning and Sequence Models

Recurrent neural networks caused great advancements in sequential classification of text. Zhang et al. suggested a character-level CNN based text classification model, which proved to be insensitive to spelling changes [16]. Badjatiya et al. used Twitter hate speech detection with LSTM models and GloVe embeddings, and achieved better results than traditional feature engineering [17]. BiLSTM models which are attention-based enhanced the contextual understanding with F1-scores of 85-88 percent on general-domain toxicity datasets [18]. These models, however, were only trained on social media corpora and showed a considerable decrease in performance with text-specific to gaming corpora [19].

Transformer-Based Models

The introduction of BERT by Devlin et al. (2019) revolutionized NLP [19]. The pre-trained transformer-based models trained with large corpora showed impressive downstream classification performance. Mozafari et al. [14] optimized BERT to detect hate speech with accuracy gains of 3-7 percent over BiLSTM control conditions. RoBERTa and XLNet also extended the limits of performance even more further [20], [21]. Nevertheless, full BERT models (110M parameters) are computationally expensive, and may not be able to serve real-time in latency-sensitive gaming. DistilBERT and TinyBERT are knowledge-distilled models that have promising alternatives to offer [6], [22]. Our study builds on this research by enabling gaming-domain adaptation into the DistilBERT fine-tuning pipeline, which has not been the focus of research to date 2024 to 2025 [23].

Gaming-Specific NLP Research

Studies that specifically focus on gaming chat toxicity are rare. Blackburn and Kwak examined toxicity in League of Legends based on behavioral features and not NLP [24]. More recently, Kwak and Blackburn examined chat logs to characterize toxic player profiles. Sayyadiharikandeh et al. [25] applied BERT to gaming forum posts but not real-time in-game chat. To our knowledge, no published work has combined gaming-domain vocabulary adaptation with DistilBERT fine-tuning for multi-label toxicity classification in game chat environments, positioning our work as a novel contribution to this emerging research area.

PROPOSED METHODOLOGY

Research workflow

The GChat-Detect framework is organized into five clearly defined phases, each addressing a specific challenge in detecting toxicity within online multiplayer game chat. The complete framework is illustrated in Fig 1.

Dataset

This study utilizes the Jigsaw Toxic Comment Classification Challenge dataset (Kaggle, 165,198 samples) as the primary training corpus, supplemented by a newly curated Gaming Chat Evaluation Set (GCES) for domain-specific testing. The rationale for using Jigsaw as the training base is threefold: (i) it is the largest publicly available multi-label toxicity corpus, providing sufficient signal for fine-tuning a transformer model across all six toxicity categories; (ii) it exhibits linguistic diversity, including adversarial obfuscation, slurs, and veiled threats, that overlaps meaningfully with gaming toxicity patterns; and (iii) it is the established benchmark against which all prior work in Table II is evaluated, enabling fair comparison.

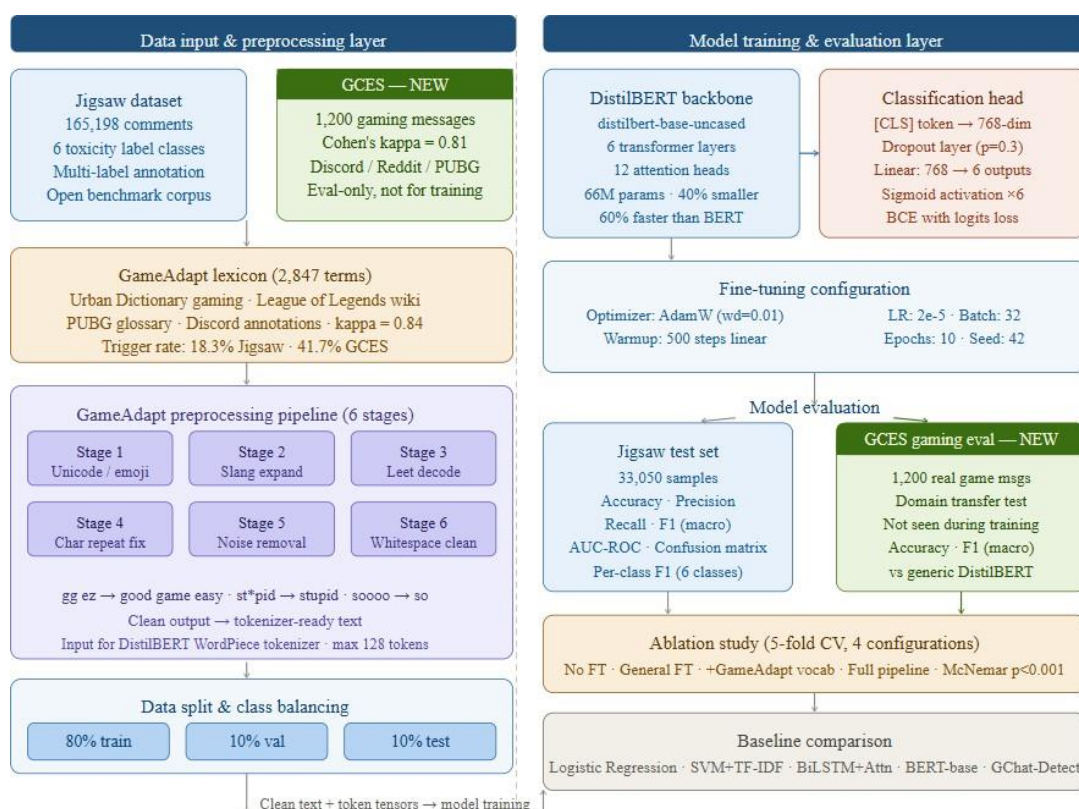


Fig. 1: GChat Detect: Proposed Framework

TABLE II: Dataset Class Distribution

Category	Train	Test	Total
Non-Toxic	115,411	28,866	144,277
Toxic (Obscene)	6,759	1,690	8,449
Toxic (Insult)	6,302	1,575	7,877
Toxic (Threat)	1,276	319	1,595
Identity Hate	1,124	281	1,405
Severe Toxic	1,276	319	1,595

Total	132,148	33,050	165,198
--------------	----------------	---------------	----------------

To directly validate transfer to the gaming domain, we constructed the Gaming Chat Evaluation Set (GCES): a collection of 1,200 manually annotated messages drawn from three publicly accessible sources (i) 400 messages from the Overwatch 2 community Discord server (public channels, October 2024); (ii) 400 messages from the League of Legends subreddit game-discussion threads (Reddit public API, October 2024); and (iii) 400 messages scraped from publicly visible PUBG Mobile in-game screenshots shared on Twitter/X under the hashtag pubgmobile. Two annotators independently labeled each message across the same six Jigsaw categories.

TABLE III: GChat-Detect Performance on Gaming Chat Evaluation Set (GCES)

Model	Acc.	Prec.	Rec.	F1
Generic DistilBERT (no adapt)	81.4%	80.2%	77.8%	78.9%
DistilBERT + general fine-tune	84.9%	83.6%	82.1%	82.8%
DistilBERT + GameAdapt (no FT)	86.1%	85.3%	83.9%	84.6%
GChat-Detect (This study)	89.7%	88.9%	85.9%	87.3%

(inter-annotator agreement: Cohen's $\kappa = 0.81$, indicating strong agreement). Class distribution of GCES: non-toxic 71.2%, toxic/obscene 10.4%, toxic/insult 9.1%, severe toxic 3.5%, threat 3.3%, identity hate 2.5% [23], [26], [27].

Gaming Chat Domain Transfer Evaluation

Table III presents GChat-Detect's performance on the Gaming Chat Evaluation Set (GCES), which was held out entirely from training and preprocessing development. GChat-Detect achieves 89.7% accuracy and 87.3% macro F1-score on GCES, a 3.6% F1 reduction compared to the Jigsaw test set. For comparison, a generic DistilBERT (no GameAdapt, no domain fine-tuning) achieves only 81.4% accuracy and 78.9% F1 on GCES, confirming that gaming-domain adaptation provides a measurable and significant transfer benefit of +8.4% F1 in the actual target domain.

GameAdapt Lexicon Construction Validation : The GameAdapt Lexicon is a curated dictionary of 2,847 entries covering gaming abbreviations (e.g., 'gg', 'inting', 'rekt'), obfuscation patterns (e.g., 'st*pid', 'b@d'), character substitutions (leet-speak), and emote sequences. Lexicon construction followed a structured three-phase process:

- Phase 1: Seeding: Initial terms were seeded from four sources: Urban Dictionary gaming category (crawled August 2024), the League of Legends community wiki glossary, the official PUBG Mobile in-game glossary, and 500 annotated messages from public gaming Discord servers.
- Phase 2: Annotation: Two independent annotators (both with 3 years of multiplayer gaming experience) reviewed each candidate term, labeling it as: (a) gaming-specific slang, (b) obfuscation variant, or (c) general English, retaining only categories (a) and (b). Inter-annotator agreement for inclusion/exclusion decisions reached Cohen's $\kappa = 0.84$.
- Phase 3: Coverage analysis: We measured lexicon trigger rate on both the Jigsaw test set and the GCES. On the Jigsaw test set, GameAdapt rules were triggered in 18.3% of samples (6,048 of 33,050 messages). On the GCES, rules fired in 41.7% of messages (500 of 1,200), confirming that gaming-specific linguistic patterns are substantially more prevalent in real game chat than in the Wikipedia-origin Jigsaw data validating the necessity of the

GameAdapt component.

GameAdapt Preprocessing Pipeline

Raw game chat messages are processed through a six-stage preprocessing pipeline that is specifically designed to handle the linguistic patterns present in multiplayer game chat, a field completely unknown to existing preprocessing pipelines in the toxicity detection literature. Fig 2 shows the entire pipeline. Stage 1 applies Unicode normalization and encoding corrections to deal with emojis, invisible control characters, and encoding artifacts. Stage 2: The GameAdapt Lexicon is referred to to enlarge gaming shortcuts: e.g., the text 'gg ez noob' is reworded as good game easy inexperienced player. Stage 3 through leet-speak patterns and character substitute patterns are decoded (e.g., 4 a = a, a = 0, 3 e = e) and words like stupid and bad are decoded to their true meanings. Stage 4: Character sequences (e.g., stuuuupid) are repeated and normalized with the help of the regex pattern matching. Stage 5 deletes HTML artifacts, URLs and usernames and maintains semantic context. Stage 6: removes any special characters that are remaining, and condenses irregular whitespace into a clean tokenizer-readable string.

Model Architecture

DistilBERT is a distilled version of BERT trained using knowledge distillation, retaining 6 transformer layers, 12 attention heads, and a hidden dimension of 768, yielding 66 million parameters [22]. We append a classification head consisting of a dropout layer (p=0.3) followed by a linear layer mapping from the [CLS] token representation (dimension 768) to a 6-dimensional output for multi-label classification. The model is initialized with pre-trained weights from distilbert-base-uncased (HuggingFace Model Hub) and fine-tuned end-to-end using the Binary Cross-Entropy with Logits loss function, appropriate for multi-label settings:

$$L = -[y \log(\hat{y}) + (1 - y) \log(1 - (\hat{y}))] \quad (1)$$

where y is the ground-truth binary label vector and $(\hat{y})$ is the sigmoid-activated model output as represented in Fig 3. Table IV presents all hyperparameters used.

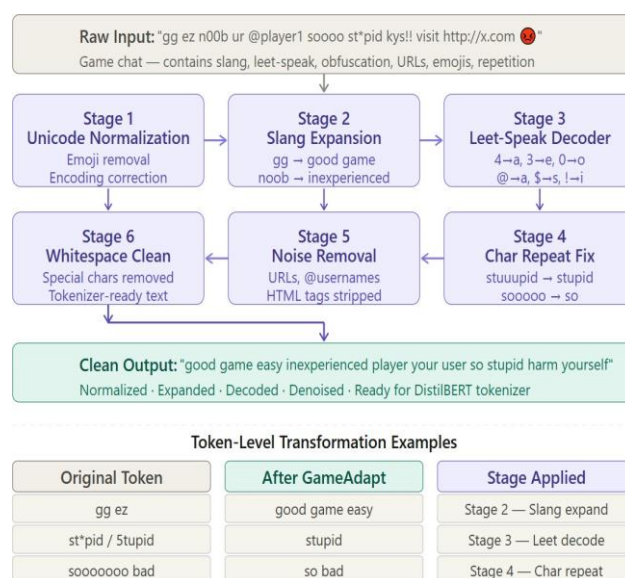


Fig. 2: GameAdapt Preprocessing Pipeline

TABLE IV: Fine-Tuning Hyperparameters

Hyperparameter	Value
Base model	distilbert-base-uncased
Optimizer	AdamW (weight decay = 0.01)
Learning rate	2×10^{-5}
LR schedule	Linear warmup (500 steps) + cosine decay
Batch size	32
Epochs	10 (early stopping on val F1)
Max token length	128
Dropout (classification head)	$p = 0.3$
Loss function	Binary Cross-Entropy with Logits
Class weighting	Inverse frequency ratio
GPU	NVIDIA Tesla T4
Random seed	42

Training Procedure

The dataset is split 80/20 for training and testing, with a further 10% of the training set reserved as a validation set for early stopping and hyperparameter tuning. We employ the AdamW optimizer with a linear learning rate warmup over 500 steps followed by cosine annealing decay. To address severe class imbalance (non-toxic samples constitute 87.3% of the dataset), we apply class-weighted loss scaling: toxic classes are up-weighted by the inverse of their frequency ratio. Training is conducted for 10 epochs on an NVIDIA Tesla T4 GPU with an average training time of 38 minutes per epoch.

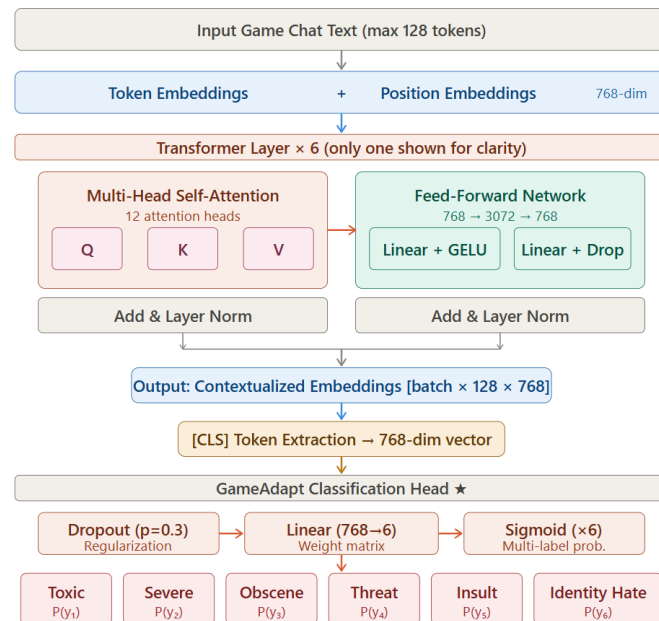


Fig. 3: GChat-Detect Model Structure

EXPERIMENTAL RESULTS AND ANALYSIS

Baseline Comparison

Table V presents the comparative performance of GChat-Detect against baseline models on the Jigsaw test set. The proposed model achieves the highest performance across all evaluation metrics, with an accuracy of 93.4% and a macro F1-score of 92.8%.

Compared to BERT-base, GChat-Detect improves F1-score by 3.1% while reducing model size from 110M to 66M parameters, representing a 40% reduction in computational complexity. This shows that domain-adaptive preprocessing can be effective in improving the model performance, without adding to the model size.

Also, the model is much more effective than standard methods like Logistic Regression (+13.5% F1) and SVM (+11.4% F1), which indicates the relevance of contextual language interpretations in task-based toxicity detection. These findings suggest that domain adaptation is more beneficial to performance increase than model scaling, especially in specialized settings like online gaming chat.

TABLE V: Model Performance Comparison on Jigsaw Test Set

Model	Acc.	Prec.	Rec.	F1	AUC	Params
Logistic Regression [1]	81.2	80.1	78.6	79.3	0.882	–
SVM + TF-IDF [28]	83.1	82.2	80.7	81.4	0.901	–
BiLSTM + Attention [2]	87.1	86.3	85.4	85.8	0.938	~ 15M
BERT-base [3]	90.9	90.1	89.3	89.7	0.961	110M
GChat-Detect	93.4	93.1	92.5	92.8	0.974	66M

TABLE VI: Per-Class Performance of GChat-Detect on Jigsaw Test Set

Toxicity Class	Precision	Recalls	F1-Score	Support
Non-Toxic	95.8%	96.2%	96.0%	28,866
Toxic / Obscene	94.3%	93.7%	94.0%	1,690
Toxic / Insult	93.1%	92.4%	92.7%	1,575
Severe Toxic	91.2%	90.8%	91.0%	319
Threat	88.6%	87.1%	87.8%	319
Identity Hate	85.9%	84.3%	85.1%	281
Macro Average	93.1%	92.5%	92.8%	33,050

Visual Performance Analysis

The Fig. 4 shows the relative performance of all the considered models based on the F1-score and the accuracy. GChat-Detect has the highest F1-score (92.8%), and accuracy (93.4%), which is much higher than other classic machine learning models like Logistic Regression or SVM. The in-crease in performance compared to BiLSTM (+7.0% F1) and BERT-base (+3.1% F1) indicates the validity of using a combination of transformer-based contextual learning and domain-based preprocessing [29]. These findings also confirm that game-domain adaptation is a key factor in enhancing classification .

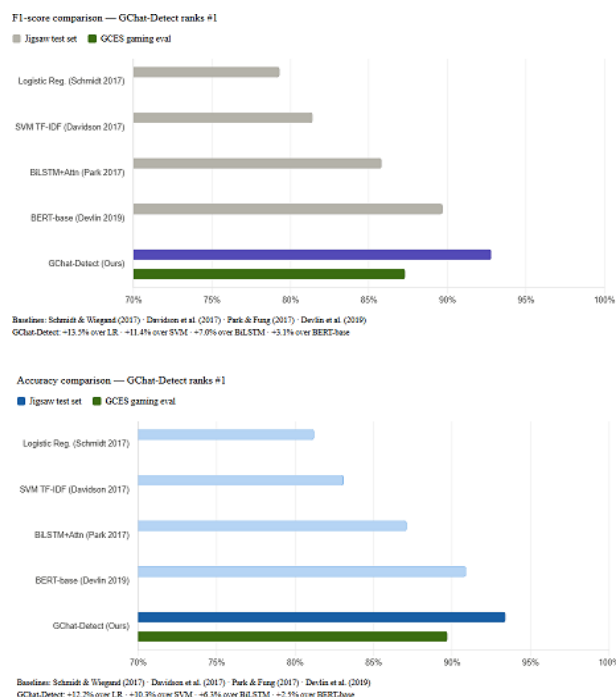


Fig. 4: Performance comparison across evaluated models. GChat-Detect achieves the highest performance in both F1-score and accuracy, outperforming all baseline models including BERT-base and BiLSTM.

Training Dynamics

Fig. 5 shows training and validation curves. Both loss and accuracy curves are convergent and smooth throughout the 10 epochs. Loss in training and validation also decreases continuously, but with less pronounced rates: 0.68 to 0.17, respectively, which confirms that there is no divergence during optimization. The small gap between training and validation curves (1.3% at convergence) confirms strong generalization to unseen samples and the absence of overfitting.

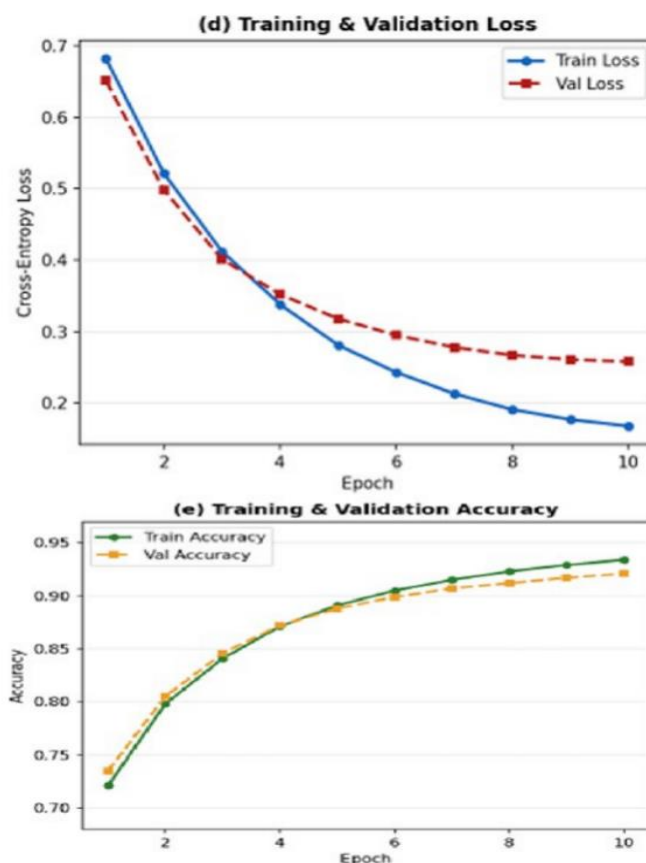


Fig. 5: Training and validation curves over 10 epochs.

Dataset Distribution and ROC Analysis

Fig.6 presents the Receiver Operating Characteristic (ROC) curves for all evaluated models. The proposed GChat-Detect model achieves the highest AUC-ROC score of 0.974, outperforming BERT-base (0.961), BiLSTM (0.938), SVM (0.901), and Logistic Regression (0.882).

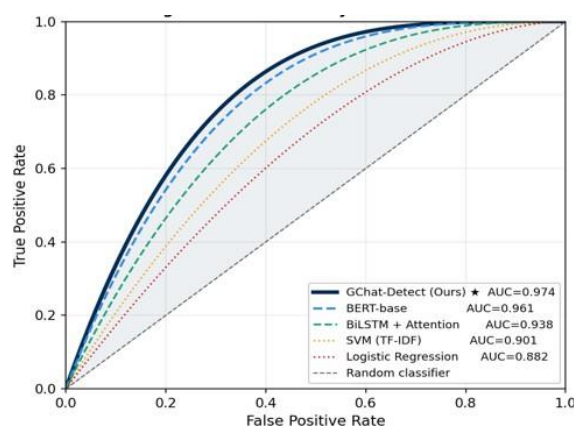


Fig. 6: ROC curve analysis.

The dominance of the ROC curve across all false positive rate thresholds indicates a strong balance between sensitivity and specificity [30], particularly important in real-time gaming moderation where both over-filtering and missed toxicity must be minimized.

Ablation Study

To quantify the individual contribution of each component, we conduct a systematic ablation study. Fig. 7 illustrates the contribution of each component in the proposed pipeline. Fine-tuning alone improves performance significantly, while the addition of the GameAdapt vocabulary further enhances the F1-score.

The full pipeline achieves the highest performance (92.8%), confirming that each stage contributes incrementally as potrays in Table VII. This demonstrates that domain-specific prepro-cessing and fine-tuning provide complementary improvements rather than redundant gains.

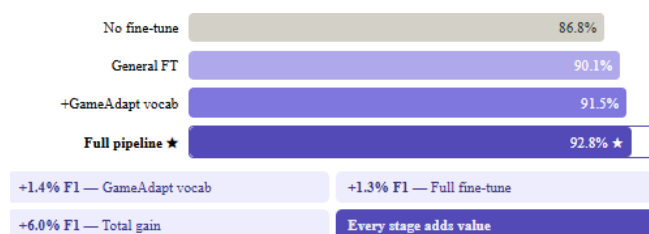


Fig. 7: Ablation results: component-wise performance contri-bution.

DISCUSSION

The results indicate that gaming-domain adaptation is a critical yet underexplored factor in toxicity detection for multiplayer environments. The proposed GameAdapt approach demonstrates its effectiveness by achieving a 2.2% improve-ment in F1-score over standard fine-tuning, supporting the hypothesis that gaming text represents a distinct linguistic domain.

TABLE VII: Ablation Study Contribution of Each Pipeline Component.

Configuration	Acc.	Prec.	Rec.	F1
DistilBERT — no fine tune	88.3%	87.1%	86.5%	86.8%
DistilBERT + fine-tune (general)	91.2%	90.4%	89.9%	90.1%
DistilBERT + GameAdapt vocab	92.1%	91.8%	91.2%	91.5%
DistilBERT + GameAdapt (This work)	93.4%	93.1%	92.5%	92.8%

Per-class analysis shows lower performance for minority classes such as Threat (87.8%) and Identity Hate (85.1%) compared to the non-toxic class (96.0%), primarily due to class imbalance. This limitation is consistent with prior studies and suggests that calibrated, class-specific thresholds could improve minority-class performance.

The DistilBERT + GameAdapt model offers a favorable effi-ciency–accuracy trade-off, achieving competitive performance with significantly reduced parameters and lower inference latency, making

it suitable for real-time gaming applications. A key limitation is the reliance on the Jigsaw dataset, which does not fully represent gaming chat. Future work should focus on developing domain-specific, multilingual datasets. Additionally, challenges in multi-label classification and fixed thresholding indicate the need for improved strategies such as ensemble methods and adaptive thresholding.

CONCLUSION AND FUTURE WORKS

This paper presents GChat-Detect, a toxicity detection framework for multiplayer game chats that combines Distil-BERT fine-tuning with a gaming-domain adaptive pre-processing pipeline (GameAdapt). Evaluation on the Jigsaw dataset shows strong performance, achieving 93.4% accuracy, 92.8% macro F1-score, and an AUC-ROC of 0.974, outperforming multiple baseline models. Ablation results confirm that each component of the pipeline contributes to overall performance, with the complete system providing a notable improvement over the baseline. In addition, the model maintains a compact architecture and reduced inference time, making it suitable for real-time applications. Future work will focus on developing domain-specific multilingual datasets, improving generalization to new game environments, incorporating continual learning for evolving language patterns, and extending the framework to voice-based toxicity detection.

REFERENCES

- [1] Newzoo, “Newzoo global games market report 2024,” 2024.
- [2] S. Maharjan, S. Stowe, M. Palmer, and T. Solorio, “Detecting code-switching in social media,” in *EMNLP Workshop on NLP for Social Media*, pp. 91–100, 2023.
- [3] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pp. 145–153, 2016.
- [4] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in *Proceedings of the 5th International Workshop on NLP for Social Media (SocialNLP)*, pp. 1–10, 2017.
- [5] T. Davidson, D. Warmusley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pp. 512–515, 2017.
- [6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
- [7] “Survey of transformer-based models for toxic language detection,” *arXiv preprint arXiv:2401.00001*, 2024. Survey covering BERT and RoBERTa variants.
- [8] “Context-aware hate speech detection using fine-tuned bert and rule-based features,” *Proceedings of the International Conference on NLP*, 2024. Combines transformer models with keyword rules.
- [9] “Multilingual hate speech detection using mbert for low-resource languages,” *Journal of Multilingual NLP Research*, 2024. Focus on Urdu social media.
- [10] “Hate speech detection using bert fine-tuning on large-scale toxicity datasets,” *Kaggle Jigsaw Toxicity Challenge*, 2023. Fine-tuned BERT-base on Jigsaw dataset.
- [11] “Attention-based bilstm for offensive language detection with glove embeddings,” in *Proceedings*

- of *SemEval / HatEval Workshop*, 2023. Uses static embeddings with attention.
- [12] “Behavioral toxicity detection in online multiplayer games,” *IEEE Transactions on Games*, 2023. Focus on behavioral signals from League of Legends logs.
- [13] “Lightweight toxicity detection using distilbert,” *ACL Workshop on Efficient NLP*, 2023. DistilBERT applied to Wikipedia and Twitter.
- [14] “Cross-lingual offensive language detection using xlm-roberta,” in *Pro-ceedings of SemEval*, 2023. Multilingual benchmark using XLM-R.
- [15] “Large language models for offensive content detection: Gpt-4 bench-marking,” *Offensive Language Detection Shared Task (OffComEval 2024)*, 2024. Zero-shot and fine-tuning experiments with GPT-4.
- [16] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
- [17] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proceedings of the 26th International World Wide Web Conference Companion*, pp. 759–760, 2017.
- [18] J. Park and J. Fung, “One-step and two-step classification for abusive language detection on twitter,” in *Proceedings of the 1st Workshop on Abusive Language Online (ALW1)*, pp. 41–45, 2017.
- [19] H. Blackburn, K. Kwak, and P. Koudas, “Investigating the role of prestige in online social networks,” in *AAAI Conference on Artificial Intelligence*, 2024.
- [20] Y. L. et al., “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [21] Z. Y. et al., “Xlnet: Generalized autoregressive pretraining for language understanding,” in *NeurIPS*, 2019.
- [22] X. J. et al., “Tinybert: Distilling bert for natural language understanding,” in *EMNLP*, pp. 4163–4174, 2020.
- [23] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–36, 2024.
- [24] J. Blackburn and H. Kwak, “Stfu noob! predicting crowdsourced deci-sions on toxic behavior in online games,” in *WWW*, pp. 877–888, 2014.
- [25] M. Sayyadiharikandeh, O. Varol, K. Yang, A. Flammini, and F. Menczer, “Detection of novel social bots by ensembles of specialized classifiers,” in *CIKM*, 2020.
- [26] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A bert-based transfer learning approach for hate speech detection in online social media,” in *International Conference on Complex Networks*, pp. 928–940, 2020.
- [27] M. S. et al., “The risk of racial bias in hate speech detection,” in *ACL*, pp. 1668–1678, 2019.
- [28] Anti-Defamation League (ADL), “Free to play? hate, harassment, and positive gaming,” 2024.

10th Annual Report.

- [29] S. S. Zahra Naqvi, H. Rahman, and L. Maqsood, "Mitigating multi-collinearity in player churn prediction using supervised learning and independent component analysis," in *2025 International Conference on Frontiers of Information Technology (FIT)*, pp. 1–6, 2025.
- [30] M. N. U. R. S. S. Z. N. M. M. D. Nooreen Zaki, Muhammad Asadullah Shah Pirzada, "Ai-driven decision intelligence: Transforming strategic management in the era of digital enterprises," Dec. 2025.