ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

PREDICTING TOXIC BEHAVIOUR IN ONLINE MULTIPLAYER GAMES: A MODEL-BASED APPROACH

Mr. Gaurav Ramchandra Tukrul[1], Dr. Siddhesh Kadam[2]
[1] *Student, Department of Data Science, Kirti. M. Doongursee College, University of Mumbai, India.*
Email: gauravtukrul399@gmail.com
[2] *Assistant Professor, Department of Computer Science, Kirti. M. Doongursee College, University of Mumbai, India.*
Email: siddhesh.kadam@despune.org

Abstract
It has been proven that anti-social behavior perpetrated by certain members of community streamers and players has consequences on the welfare of the community and its members, as well as the sustainability of the community itself. This research focuses on the survey-based data collected from the community and applies a binary machine learning model to identify the toxicity predictors as well as hyper-parameter settings. Two tasks are created: a binary classifier to predict whether a player has been exposed to toxicity and a multi-label classifier to predict the form of toxicity the player has been exposed to. Random Forest classifiers, which are weighted to mitigate class imbalance with the help of a balanced SMOTE model to mitigate class imbalance, were tested, along with Logistic Regression, Support Vector Machine, and XGBoost. The experimental finding shows that Random Forest outperformed its competitors by a wide margin, although all investigated classifiers were acceptable. Random Forest has a balanced accuracy of 73.1%, 0.715 F1 scores, and 0.788 ROC-AUC, making this one of the best classifiers for exposure risk prediction. For the multi-label classification, the Random Forest classifier produced the lowest Hamming Loss of 0.357 and the Micro- and Macro-F1 scores of approximately 0.65, which are the highest for that dataset. Support Vector Machines were predictable, Random Forest classifiers were the most stable, Logistic Regression classifiers balanced efficiency and interpretability, and Random Forest classifiers from the ensemble gained the most attention with the proven functionalities.

Keywords: Toxicity Detection, Online Multiplayer Games, Machine Learning, Multi-Label Classification.

► *Corresponding Author: Mr. Gaurav Ramchandra Tukrul*

1. Introduction:

Online multiplayer games have become complex social systems where millions of players engage and interact directly in real-time. With any work that's cooperative, competitive or community-focused, it brings an extensive amount of social issues with regard to harassment, hate speech or cheating." These problems not merely degrade playing experience, but also lead to some psychological issues for players, because gaming disengagement and divide gaming communities. The problem of toxicity in online gaming has become a longstanding concern for players and

developers, aggravated in part by the anonymity and competitive environment of the gaming industry.

To that end, many developers and platform operators have utilized a mixture of techniques for moderation, including automated content filtering and player-reported systems, to try to reduce toxicity. Each of those weigh personally for them the pros and cons, especially with the balance between each of their enforcement. Little to no enforcement means moderation freedom, which is known to encourage harmful behavior, while strict enforcement risks suppressing genuine expression. All these factors, alongside evaluating current moderation techniques, form the basis for creating safer and more inclusive gaming environments.

This study focuses on exploring patterns of toxic behavior in online gaming communities and evaluating the effectiveness of existing moderation mechanisms. By combining survey data with community managers and policymakers to strike a balance between encouraging freedom of expression and ensuring player well-being.

2. Literature Review:

McGill and Ubisoft researchers created a BERT-derived model designed to identify toxic behavior in players' interactions in Rainbow Six Siege and For Honor by using 194,000 ingame chat lines that had been annotated. The model's context-aware, history-aware capabilities considering the game's context and chat history boost the accuracy of the model to 82.95% precision and 83.56% recall, which surpasses prior models' performance in toxicity detection. [1]

While focusing on the classification of toxicity, the author also examines six deep learning models, including CNN, LSTM, and BERT, and highlights various word embeddings and their impacts. The results show that BERT is more effective than the other models, further demonstrating its capability to comprehend and detect harmful interactions in text-based communication.[2]

This research presents HateXplain, the very first hate speech dataset benchmark that incorporates classification, target community labeling, and rationales given by humans for labeling. The authors collected over 20,000 posts from Twitter and Gab, which were labeled using Amazon Mechanical Turk, and labeled as hate speech, offensive, or normal. Moreover, the targeted community was identified, and portions of the text justifying the labeling were highlighted. Analysis showed that benchmarks for human explain ability failed due to lacking alignment with human reasoning, despite the best performing models excelling on classification. Models trained with human rationales improved both the explain ability of model decisions and bias against minority groups. The dataset and code provided alongside the paper offer a resource to construct systems for hate speech detection that offer flexible bias and explain ability, and thus the paper contains less biased and interpretable systems.[3]

This work addresses the fragmentation of toxic comment datasets by introducing a unified software tool that retrieves, standardizes, and maps over 40 public datasets into a common format. To make cross-dataset training and analysis easier, the collection, which covers 12 platforms and 13 languages, unifies disparate file formats, labeling schemes, and metadata. The mapping system allows researchers to harmonize disparate toxicity-related labels to create consistent binary or multi-class setups. This integration highlights biases like keyword-triggered sampling that may restrict real-world applicability while facilitating largescale multilingual training, dataset comparison, and reproducibility. The open-source project aims to develop more reliable and broadly applicable toxic comment detection models through further development.[4]

This study presents a significant dataset of over a million in-game chat messages containing various toxicity labels, ranging from disrespectful to profane language and harassment. This

dataset establishes a reliable benchmark for machine learning toxicity assessment frameworks, aiding in the development of more sophisticated models. The dataset is also beneficial for the development of online gaming analytics and predictive models.[5]

This research aims to identify toxic behavior in online games and is published by Martens, Shen, and Iosup. It relies on server-side data such as player reports, match results, and chat logs which enable the identification of toxic players to initiate a classification framework designed for preemptive toxic behavior detection. This framework, which relies on a gameplay data and chat interactions, creates a holistic framework to predict and understand toxicity in online games and resolves the limitation of most prior studies concerning gameplay data utilization. [6]

In total, 3 major online platforms CNN, Breitbart, and IGN, and their permanently suspended users were investigated in depth and over time. Facebook banned users were less readable and less positive, and engaged with more posts, but in a smaller number of threads than their never-banned counterparts. FBUs' posts were and remained of lower community quality, refinement, and collaboration and were even contributing further, receptively, on a community level, to their expedited bans. However, a range of antisocial behaviors were explored and the effects of present-day and future under-censorship disparate antisocial behaviors were examined. With only 5 to 10 posts, the authors constructed a system that monitored the changes in users' community integration and the level of unsanctioned antisocial behaviors during a given time to predict with an accuracy of over 80% a permanent ban. The authors of the paper aim to present the findings in a way that can promote healthy online conversations, to enable the system to determine expeditious antisocial engagement and suggest early intervention to the community.[7]

This research analyzes how Naive Bayes, SVM, and Random Forest, with the aid of TF-IDF and other features, detect toxicity. Even though the architectures of less democratic, shallow learning, these models are still inferior to deep learning, and for more reasonable counterarguments. Nevertheless, the investigation remains relevant for contextual and historical comparison with other models for toxicity classification. [8]

Ubisoft and McGill researchers applied a Small LLM (GPT-4o-mini) model to soft prompting. Using the ToxBuster framework, they produced a model that accurately and efficiently detected toxic comments in a range of 15 datasets across 7 languages in a resource-lean multilingual environment. [9]

Praise-based behavioral psychology applications in identifying and fostering the respect and recognition of positive behaviors in online gaming communities are the center of this study. The introduction of positive preemptive messages, messages of community standards, and systems of rewards. The community behavioral nudges that are designed to focus on the respectful interactions show clear gaming etiquette improvements. [10]

3. Objective:

1. To conduct a comparative analysis of machine learning models to detect and predict toxic behaviour in online gaming platforms.
2. To design or evaluate predictive models that can proactively detect and mitigate toxicity in real-time gameplay environments.

4. Methodology:

We suggest the method in used in this work to be systematic behavior identification and prediction in toxic behavior prediction in online multi-player games. This includes datasets preparation and preprocessing, modeling, hyperparameter tuning, evaluation and comparisons.

4.1. Dataset Collection:

The first source of data was obtained through a structured survey on player experiences in online multiplayer games. The survey responses contained information on demographics, game preferences, frequency of toxicity and nature of the encountered toxic behavior, as well as reporting/moderation decisions.

4.2. Data Preprocessing:

In order to guarantee the quality and consistency of data, the following two actions were taken.

- Label Encoding and Mapping: Binary or categorical outcomes were encoded/converted into numerical one (e.g., Encountered Toxicity → {Yes:1, No:0}, Reporting Outcome → {banned=1, else=0}).
- Multi-label Encoding: For multi response questions (e.g., Types of Toxicity, Toxicity Impact), we utilized the Multi Label Binarizer to transform several text categories into binary vectors.
- Frequency Mapping: Toxicity Frequency responses were mapped on a numerical scale of 1–5.
- Feature Engineering: Moderation Gap (the gap between the moderation effectiveness and frequency scores, capturing disparity in managing toxicity) is a new variable introduced to account for this.
- Treatment of Missing Data: Categorical missing data were imputed with "None" and numerical missing values were imputed with 0.
- Feature Encoding: The age group and game categories were one-hot encoded to develop features ready for modeling.

4.3. Problem Formulation:

Two classification tasks were defined:

- Binary Classification (Exposure Risk):– Predict whether a player is exposed to toxicity or not (Yes/No).
- Multi-Label Classification (Type of toxic): Detecting one or more type/ category of toxicity encountered by players.

4.4. Data Splitting and Balancing:

Stratified sampling was used to divide the dataset into training (80%) and testing (20%) datasets. In order to overcome such class imbalance, we used SMOTE (Synthetic Minority Oversampling Technique) on the training data for binary classification by balancing the exposure risk classes.

4.5. Algorithm Selection and Justification:

We chose four machine learning algorithms in this study, including Logistic Regression (SVM with Kernel), Random Forest (RF) and XGBoost. The reasons for selecting these algorithms are as follows:

- Logistic Regression: We chose logistic regression as a simple and interpretable baseline for binary classification tasks so we could establish some kind of benchmark.
- SVM (SVM-Kernel): SVM was chosen, as it can model non-linear relationships and find optimal separating hyperplanes in more complex data.
- Random Forest: Selected for its ensemble learning method which minimizes overfitting, successfully captures intricate interactions among different features, and delivers insights in feature importance.
- XGBoost: Chosen for its boosting framework, excellent performance and generalization of features, and built-in regularization.

4.6. Hyperparameter Tuning:

Hyperparameter tuning was done using GridSearchCV with 3 fold cross-validation with:

- Logistic Regression: c values (0.1

- ,1,10)
- SVM: C values (0.1, 1, 10) and kernels (linear RBF)
- Random Forest: n_estimators (100, 200) and max_depth (None, 10, 20)
- XGBoost: n_estimators (100, 200), learning_rate (0.05, 0.1), max_depth (3, 5)

4.7. Model Evaluation:

- Binary Classification (Exposure Risk): Models were evaluated using Accuracy, Precision, Recall F1 Score and ROC-AUC.
- Multi-Label Classification (Types of Toxicity): Average of Hamming Loss, Micro F1 Score and Macro F1 Score was done for evaluating the models.

We reported our results on tables, and illustrated each metric with bar charts.

4.8. Comparative Analysis:

We compared the performance of the four different algorithms. We plotted the binary and multi-label results using Matplotlib and Seaborn for improved interpretability. This framework helped us identify the best model for predicting toxicity in online gaming communities.

5. Results and Discussion:

5.1. Binary Classification Results (Exposure Risk Prediction)

The objective of the binary classification task was to anticipate whether a gamer would likely undergo toxic interactions during online play. To address the issue of class imbalance, a balanced dataset was prepared using the SMOTE technique, on which Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost, the four supervised models, were trained and evaluated.

The models were assessed on the five different metrics: Accuracy, Precision, Recall, F1Score, and ROC-AUC. These are reflected in the figures.

- Random Forest outperformed all other models with a 0.73 accuracy and with precision of 0.75, recall of 0.69, and ROC-AUC of 0.79.
- SVM equally displayed competitive performance with 0.68 accuracy and 0.73 ROCAUC, which indicates the model's capability of robust performance in capturing non-linear decision boundaries.
- Logistic Regression, with an accuracy of 0.62, displayed the interpretability of model results and a balance of precision and recall.
- XGBoost, with a 0.55 accuracy (moderate generalization), indicates that the boosting framework's general performance on self-optimizing from survey data is a possible reason for underperformance on this dataset.

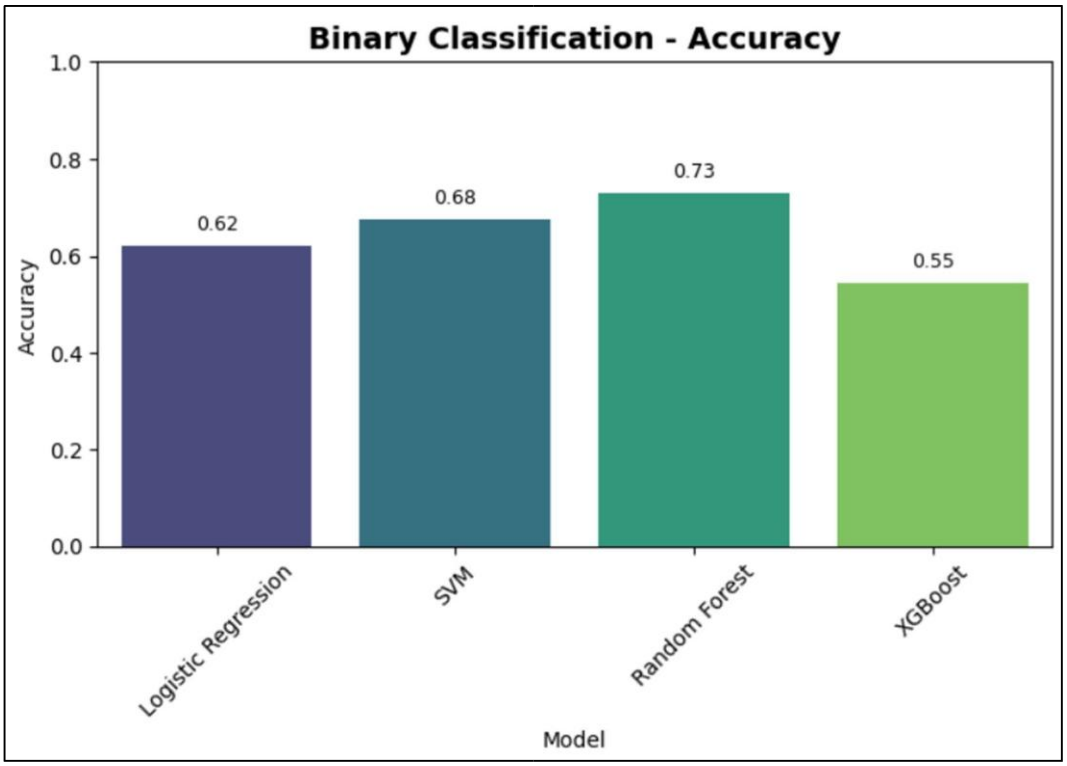

Figure 1, Binary Classification - Accuracy

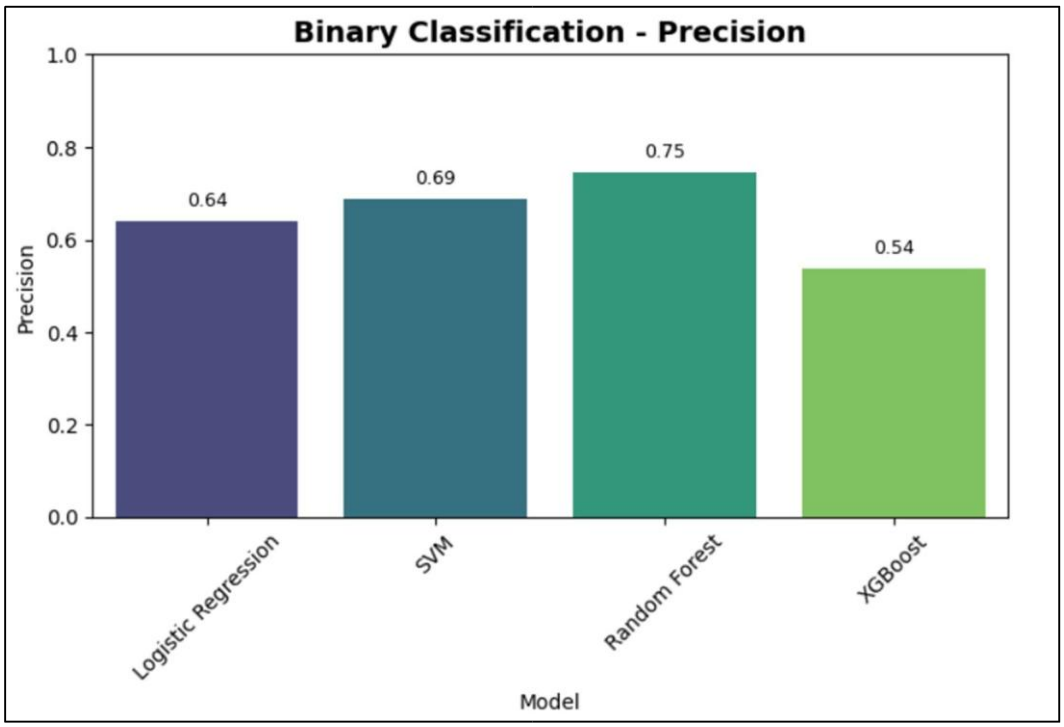

Figure 2, Binary Classification - Precision

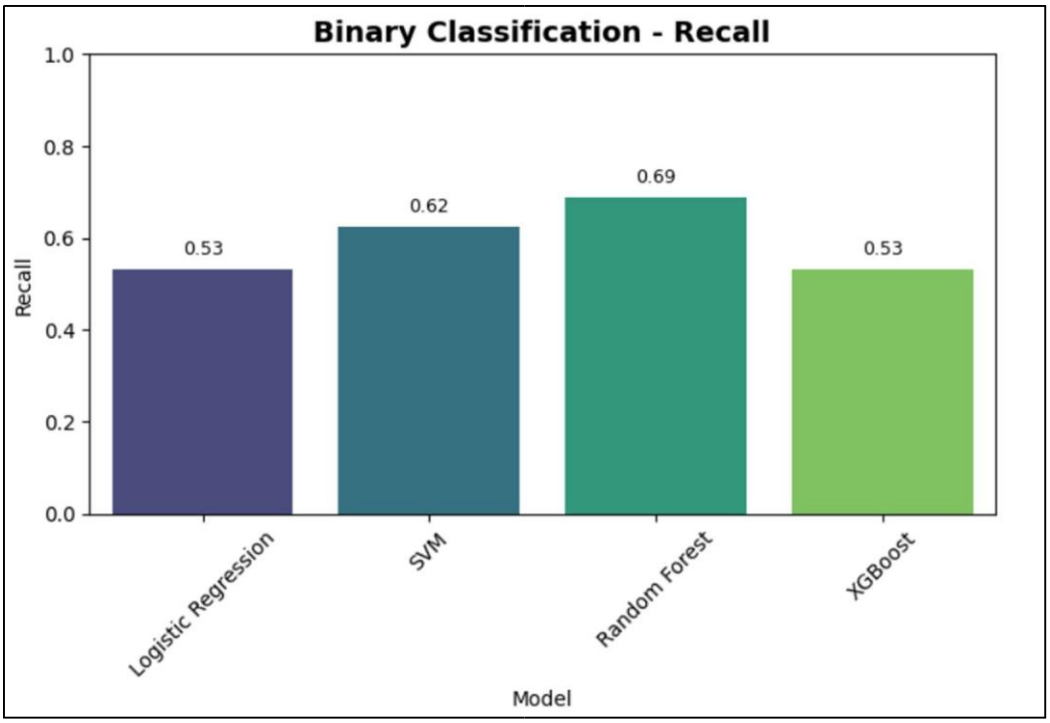

Figure 3, Binary Classification - Recall

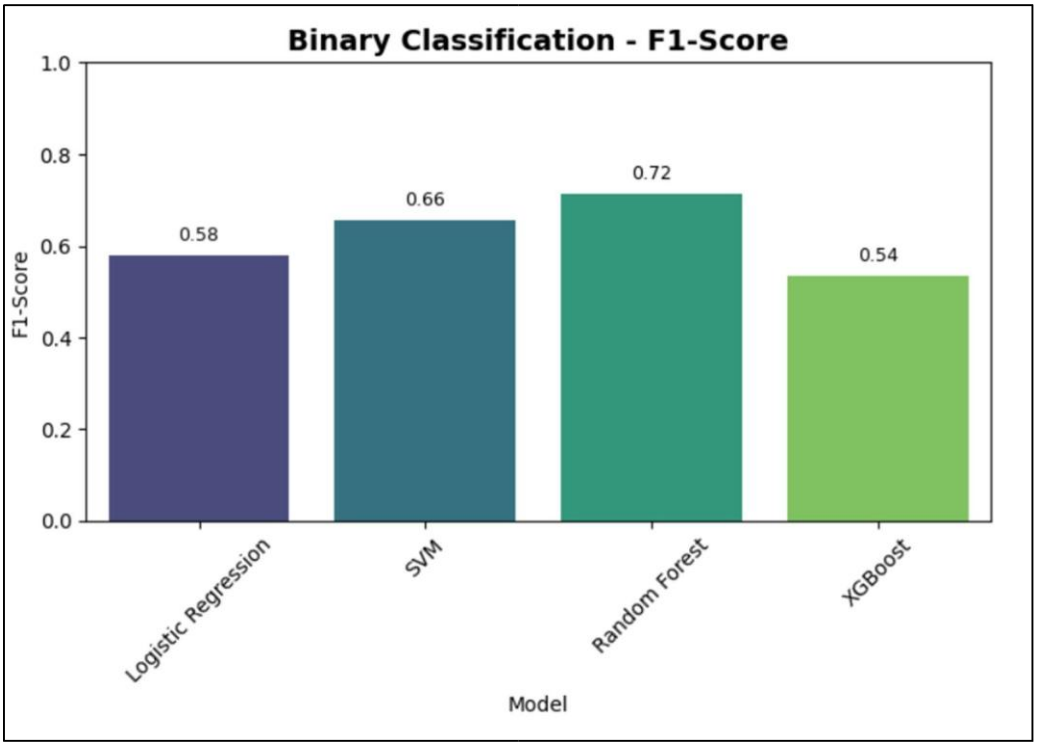

Figure 4, Binary Classification – F1-Score

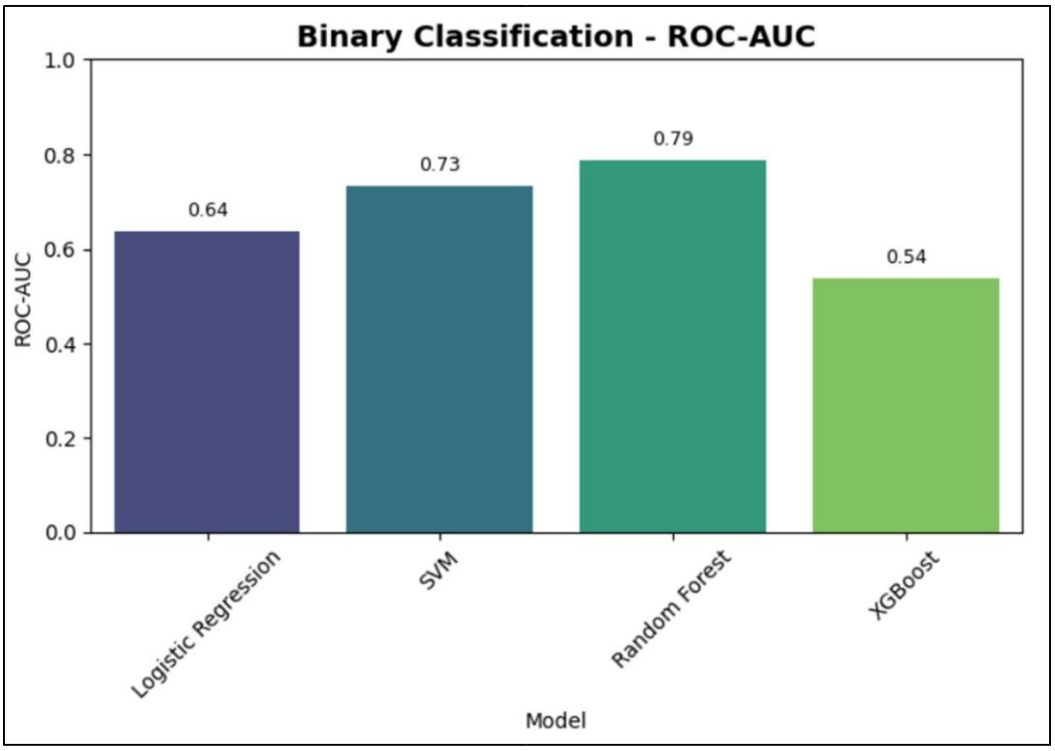

Figure 5, Binary Classification – ROC-AUC

To conclude, Random Forest was most effectively predicted the exposure risk, which suggests that ensemble model techniques can best serve high multi-faceted behavioral datasets.

5.2. Results for Multi-Label Classification (Type of Toxicity Prediction):

Hamming Loss, Micro-F1, and Macro-F1 metrics were applied for assessing the models for predicting several types of toxic activities, such as harassment, hate speech, and cheating.

- Once again, the Random Forest model displayed the most consistent performance, achieving the lowest Hamming Loss (0.36) and the highest Micro-F1 (0.65) and Macro-F1 (0.65) scores.
- With Micro-F1 at approximately 0.58–0.59, SVM and Logistic Regression provided mediocre results, suggesting an ability to model overlapping toxicity categories.
- XGBoost also had consistent results with Micro-F1 x = 0.60, affirming the consistency across various toxicity types.

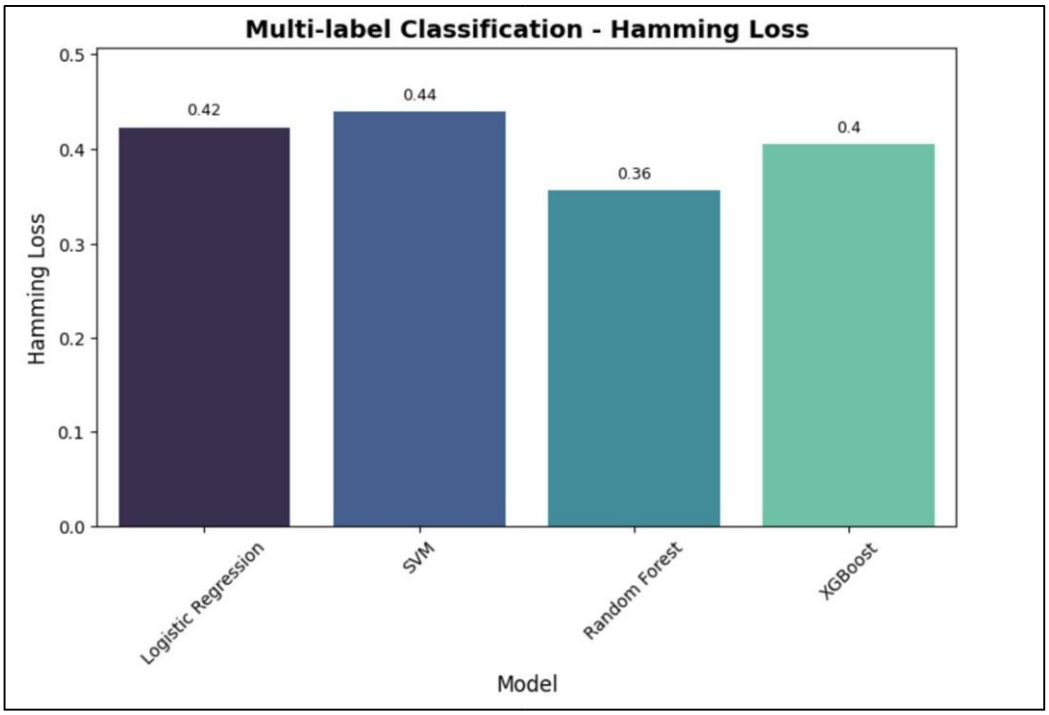

Figure 6, Multi-Label Classification – Hamming Loss

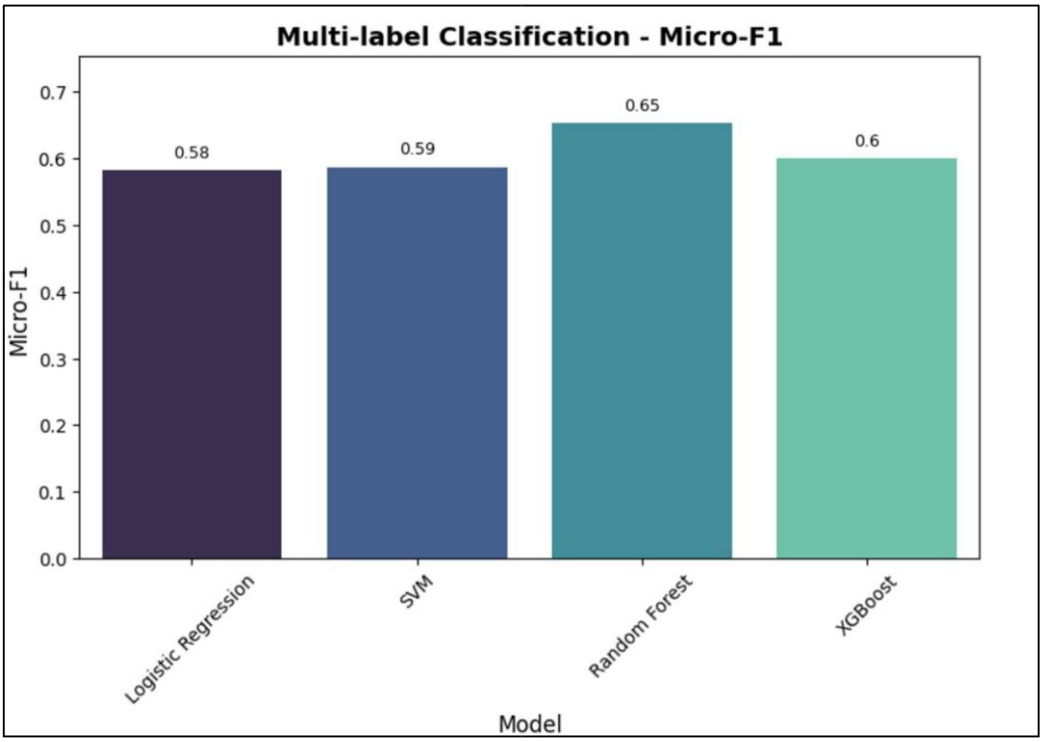

Figure 7, Multi-Label Classification – Micro-F1

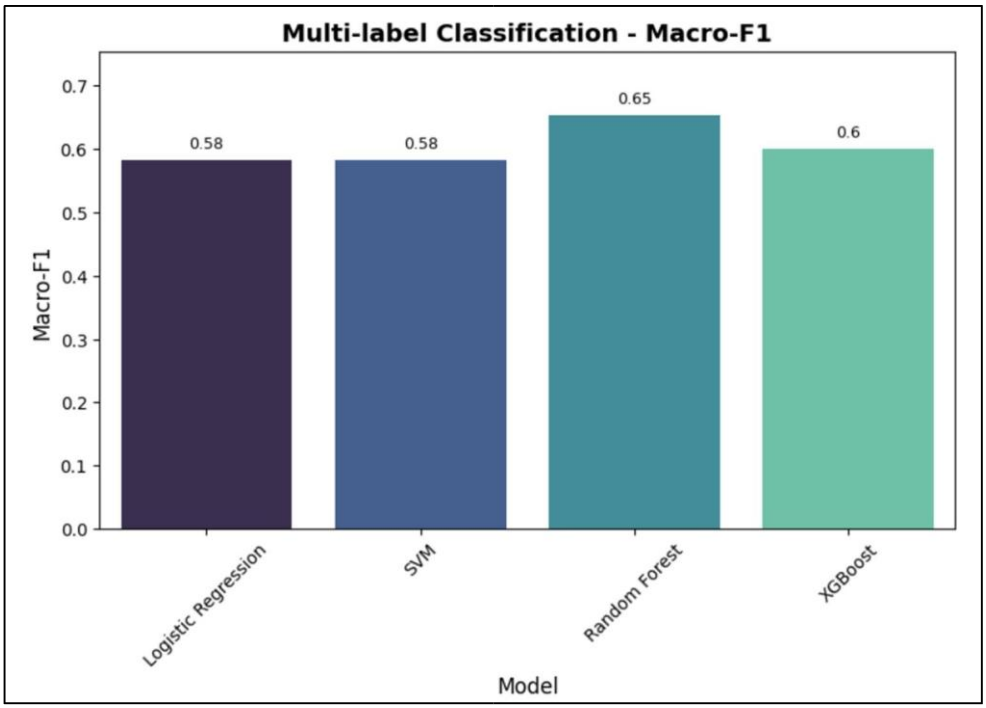

Figure 8, Multi-Label Classification – Macro-F1

These results suggest the ensemble models, especially the Random Forest model, effectively captures interdependencies within various toxic behaviors, and the linear models, while including less computational power, still provides usability and straightforward interpretability.

5.3. Comparative Discussion:

When examining both binary and multi-label classification tasks, Random Forest consistently provided the best results, exceeding 70 per cent accuracy on predicting exposure risk and generating the best F1 and ROC-AUC scores compared to all other models.

This indicates that, when looking at complex, high dimensional behavioral patterns, treebased ensembles perform better than linear and kernel-based SVM models. SVM showed strong recall, while the other two models, Logistic Regression and Lasso, stuck to the tasks and provided consistent results across all the models.

This indicates the high accuracy and high generalizability of ensemble learners when combined with feature rich survey data, creating a valuable tool to help game developers identify toxic players and better control toxicity in the game.

5.4. Findings Overview:

Random Forest outperformed competing models by scoring an accuracy of 73.1%, an F1 of 0.715, and a ROC- AUC of 0.788 in risk exposure which made it top all models in the binary classification. In addition, it also performed well in multi-label classification having Hamming Loss 0.357 which is the lowest and F1 scores of ~0.65 which demonstrates its capacity for capturing the multi relevancy nature of the different types of toxicity. This confirms the high ability of the model to approximate the complex non-linear relationships between the behavior and the demographics, thus it is the most appropriate model for estimating the presence and the type of toxicity in online games.

Due to being able to function reasonably well on moderate complexity and non-linear decision boundaries, the SVM achieved solid and consistent performances. Although slightly less successful, the lightweight nature of Logistic Regression meant it performed better in use case situations involving real time moderation and early detection automation systems. Conversely, of the other ensemble methods within the provided context, XGBoost performed the worst. This shows that more sophisticated boosting techniques may need more data and/or additional heterogeneity in features to fully unlock their benefits. Overall, the findings indicate that ensemble-based models, and more specifically the Random Forest, that gives the best trade-off between accuracy and robustness, interpretability and hence the most dependable construct for toxicity prediction in multiplayer online games.

6. Conclusion and Future Work:

Conclusion:

To sum up, the following study evaluated four machine learning models – Logistic Regression, SVM, Random Forest, and XGBoost – in order to identify and forecast toxic behavior in online multiplayer video games. According to the experimental outcomes, Random Forest was the top performer, with 73.1 percent accuracy, 0.75 precision, 0.69 recall, and 0.79 ROC-AUC when running binary classification. Besides, it also outperformed in the multi-label classification by having a low Hamming Loss of 0.36 and the largest F1 score, approximately 0.65.

The SVM model closely followed, suggesting high generalization and recall, whereas the Logistic Regression yielded a computationally efficient and interpretable baseline. XGBoost's performance was relatively weaker, which poses the possibility for simpler ensembles to outperform the boosting methods when built on structured behavioral data. Therefore, according to the findings, the ensemble learning methods, and Random Forest in particular, are the most appropriate for toxicity prediction in real-world gaming, providing interpretability, precision, and robustness.

Future Work:

To further improve prediction accuracy and real-world applications, consider the following.

- **Deep Learning Models:**

Utilize LSTM, CNN, or Transformer-based models to gain contextual and temporal understanding of player interactions.

- **Feature Optimization:**

Advanced feature selection methods like PCA, recursive feature elimination, or autoencoders could be implemented to minimize noise and attain better generalization.

- **Explainable AI (XAI):**

To ethically provide AI in gaming, applying SHAP and LIME methods will be crucial to explain the prediction of a model.

- **Cross-Platform and Multimodal Datasets:**

To enhance the model and decrease bias, use large-scale, multilingual, and multimodal datasets (text, chat logs, and audio).

- **Hybrid Approaches:**

Accuracy and interpretability can be enhanced beyond 75% by integration of Random Forest and deep learning within stacked or ensemble architecture.

7. Reference:

1. Abadji, J., Bojanowski, P., Tixier, A. J. P., & Labeau, M. (2023). ToxBuster: In-game Chat Toxicity Buster with BERT. Proceedings of the 2023 Conference on Empirical

Methods in Natural Language Processing (EMNLP). https://arxiv.org/abs/2310.18330
2. Bhatti, H., Patel, V., & Shah, K. (2023). An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection. Social Network Analysis and Mining, 13(1). https://doi.org/10.1007/s13278-023-00956-4
3. Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Antisocial Behavior in Online Discussion Communities. Proceedings of the 26th International Conference on World Wide Web (WWW), 1361–1372. https://doi.org/10.1145/3038912.3052674
4. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 12(1), 491–500.
5. Jain, A., Bansal, R., & Goyal, P. (2025). GameTox: A Benchmark Dataset for Game Chat Toxicity. Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). https://aclanthology.org/2025.naacl-short.37
6. Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015). Toxicity Detection in Multiplayer Online Games. Proceedings of the 2015 International Workshop on Network and Systems Support for Games (NetGames). https://doi.org/10.1109/NetGames.2015.7382999
7. Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745
8. Thakur, A., & Pandey, M. (2022). Analysis of Online Toxicity Detection Using Machine Learning Techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 8(2), 26–34. https://doi.org/10.32628/CSEIT228213
9. Wang, C., Jiang, Y., Abadji, J., Desaulniers, L., & Labeau, M. (2025). Unified Game Moderation: Soft-Prompting and LLM-Assisted Label Transfer. arXiv preprint. https://arxiv.org/abs/2506.06347
10. Wulf, T., & Schneider, F. M. (2024). Positive Behaviour Interventions in Online Gaming: A Review of Design Strategies and Psychological Theories. Computers in Human Behavior Reports, 10, 100097. https://doi.org/10.1016/j.chbr.2023.100097