STM JOURNALS

Trends in
Opto-electro & Optical Communication

ISSN: 2231-0401 (Online)
ISSN: 2347-9957 (Print)
Volume 15, Issue 3, 2025
September–December
DOI (Journal): 10.37591/TOEOC

https://journals.stmjournals.com/toeoc

Review TOEOC

Gradient Boosted Regression Tree Approach to Predicting Toxic Interactions on X and YouTube

Senthil P.[1], Aniruthya A.[2], Harini S.[2,*], Rahaswedha K.[2]

Abstract

In the digital age, social media platforms play a vital role in facilitating user engagement, encompassing both positive interactions and avenues for negative, often harmful behaviors. Recognizing and addressing toxic exchanges is paramount to nurturing healthy online communities and preserving users' well-being. This study introduces a novel method for identifying toxic interactions by utilizing Gradient Boosting Regression Trees (GBRT) algorithm, a machine learning approach renowned for its exceptional accuracy and ability to handle intricate, non-linear data relationships. The proposed GBRT compares five traditional classification techniques, such as Logistic Regression (LR), Random Forests (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and SGD Classifier (Stochastic Gradient Descent) which are commonly employed in toxicity identification endeavors. The comparative analysis is done using metrics like accuracy, precision, recall, and F1-score and the results show that the GBRT outperforms other compared algorithms with its overall performance. Respectively, the precision rates of GBRT, SVM, RF, LR, NB and SGD Classifier are 96, 94, 89, 88, 85, and 81%; accuracy rates of GBRT, RF, SVM, LR, NB and SGD Classifier are 95, 93, 89, 83, 80, and 78%; recall rates of GBRT, RF, SVM, NB, LR and SGD Classifier are 95, 92, 90, 87, 84, and 81%; F1-scores of GBRT, RF, SVM, LR, NB and SGD Classifier are 94, 91, 89, 86, 83, and 80%. The outcomes are achieved by conducting extensive trials on publicly available social media datasets, such as the Final Balanced Dataset and Youtoxic with the size of 57746.

Keywords: Toxic comment detection, X, YouTube, gradient boosting regression trees, logistic regression, random forests, support vector machine, Naive Bayes and SGD classifier, machine learning

INTRODUCTION

Communication is facilitated by social media sites like YouTube and Twitter, but they are sometimes tainted by hate speech, bullying, and inflammatory comments that hurt people and diminish online discourse. Effective moderation is difficult due to traditional filtering's inability to handle the intricacy of toxic language. A solution is provided by Gradient Boosting Regression Trees (GBRT), which analyze large datasets to find minute patterns of toxicity, such as contextual cues and linguistic subtleties. In order to improve content moderation and mitigate the detrimental effects of toxic behaviors on social media, the system makes use of GBRT's capacity to manage intricate, non-linear interactions.

Figure 1 shows social media interactions in two different styles, possibly from sites like Twitter or YouTube. Posts or comments with user IDs, timestamps, and text are shown chronologically on the

***Author for Correspondence**
Harini S.
E-mail: harinisaravanakumar08@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering, Karpagam College of Engineering, Coimbatore Tamil Nadu, India.
[2]Student, Department of Computer Science and Engineering, Karpagam College of Engineering, Coimbatore Tamil Nadu, India

Received Date: June 14, 2025
Accepted Date: June 19, 2025
Published Date: September 10, 2025

Citation: Senthil P., Aniruthya A., Harini S., Rahaswedha K. Gradient Boosted Regression Tree Approach to Predicting Toxic Interactions on X and YouTube. Trends in Opto-electro & Optical Communication. 2025; 15(3): 7–14p.

left. The original post is represented by the root node of the tree structure on the right, while the branching nodes display the responses. Red nodes indicate unfavorable or flagged interactions, whilst green nodes indicate neutral or good exchanges. This graphic facilitates the analysis of online discussion dynamics, sentiment, and hierarchy.

The key contributions of this work are summarized as follows: In order to increase the accuracy of content classification, this research presents a unique technique for identifying harmful interactions on social media using the Gradient Boost Regression Trees (GBRT) model. The efficacy of GBRT in content moderation is increased by its capacity to manage complicated linguistic patterns and non-linear interactions.

LITERATURE SURVEY

Online platforms have now become the mainstay of communication but often provide a cover for harmful behavior, making NLP crucial in the detection and mitigation of such content [1]. Techniques like Ridge Regression, CatBoost, and BERT have been used to model and detect varying levels of offensiveness in social media comments [2]. Malicious users hide their toxic content by embedding the text within an image. Thus, models must be able to analyze both textual and visual content to efficiently address toxic behavior [3]. Toxic comments on social networking sites often derail meaningful conversations and require techniques that can identify and mitigate specific toxic intervals within user comments [4]. Forums and online communities, while promoting idea exchange, also harbor negativity, with toxic comments fueling division and cyberbullying [5].

The effectiveness of optimizers like ADAM in training RNN architectures (LSTM, Bi-LSTM, and GRU) has been demonstrated, achieving high test accuracy (95.33%) in toxic comment classification [6]. Despite advancements, detecting harmful content remains challenging, requiring innovative frameworks to combat the proliferation of toxic texts online [7]. There are reliability issues with current moderation systems and opportunities for better solutions, with advancements in NLP and cloud computing [8]. Machine learning models have had the criticism of perpetuating biases from training datasets and inadvertently disadvantages vulnerable groups. Biases need to be addressed in order that content moderation may be fair and effective [9]. Comparative research between feature extraction techniques such as Bag of Words (BoW) and TF-IDF with deep learning methods (CNN, LSTM, and BERT) on Thai Twitter datasets have proven that pretrained models are more superior in the detection of toxicity [10].

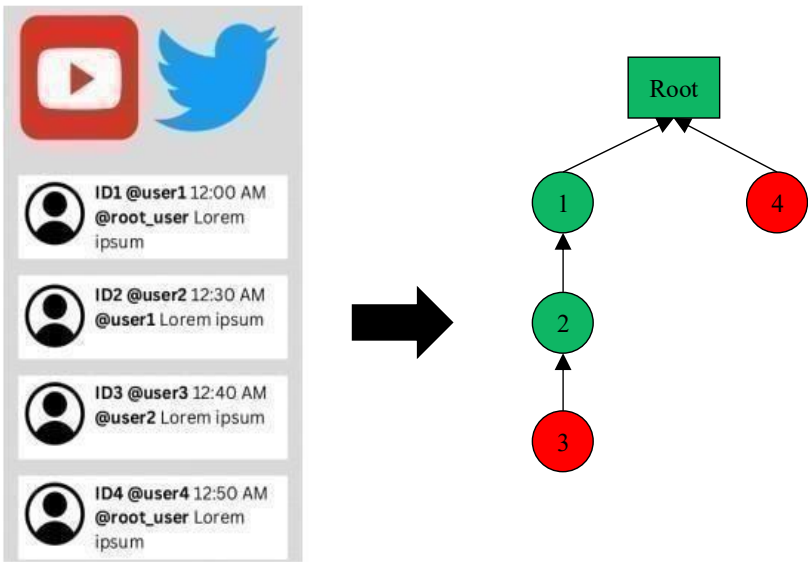

Figure 1. Shows Graphical representation of a conversation tree on YouTube and Twitter.

Multi-task learning frameworks have also proven promising in the classification of subtypes of abusive content, including aggression and personal attacks, through the exploitation of shared knowledge across tasks [11]. Single-task learning models are very heavy in terms of data and also require large computer resources, but they cannot achieve scalability to appropriately handle toxic contents [12]. NLP techniques have successfully been used at Reddit to differ between positive and negative comments by spotting cyberbullying and insults [13]. One more study that experimented with 159,000 comment dataset compares classic and deep-learning models to find whether advanced models yield better performance or not in regard to detecting the toxicity of this content [14]. Prompt optimization with large language models has enhanced moderation capabilities, especially in specialized environments such as gaming communities [15].

Hybrid models combining CNN and LSTM have been developed to detect toxic comments more effectively, especially in the case of anonymous abusive behavior online [16]. Pretrained models such as RoBERTa, which optimize hyperparameters during training to improve on BERT, have demonstrated better performance in toxicity classification [17]. The emergence of cyberbullying among teenagers via toxic comments and objectionable images indicates the requirement of models that classify multiple categories of toxicity with various degrees of accuracy [18]. Manual detection of toxic content is both time-consuming and error-prone, thus the need for an automated solution to facilitate moderation [19]. Finally, the increased ease of access of social media has amplified the spread of harassment, making scalable detection frameworks based on automation absolutely necessary for dealing with the new challenges that toxic online behavior poses [20].

MATERIALS AND METHODS

Datasets

The study uses a publicly available dataset of 57,746 social media interactions from YouTube and Twitter, labeled as toxic or non-toxic. It is divided into training (80%) and testing (20%) sets for binary classification. The dataset captures diverse language styles and contexts to ensure model robustness. Data quality improvements through preprocessing: text cleaning, tokenization, stop-word removal, and normalization feature extraction techniques including n-grams, sentiment analysis, and keyword frequency analysis GBRT was used to handle the problem as it handles the non-linear relationships of data very effectively in performing content moderation tasks.

GBRT for Toxicity Detection

A key component of the system is the Gradient Boosting Regression Trees (GBRT) model, which is intended to detect intricate toxicity patterns in social media data, including hate speech and abusive language. GBRT uses several decision trees to iteratively improve its predictions, picking up on minute details like sarcasm and context-specific harmful factors. For real-time adaptation to new toxicity trends, it necessitates additional techniques. This is addressed by the system's dynamic feedback mechanism, which takes user and moderator input into account and enables the model to adapt to novel hazardous behaviors. Figure 2 illustrates the GBRT based toxicity detection system workflow [21].

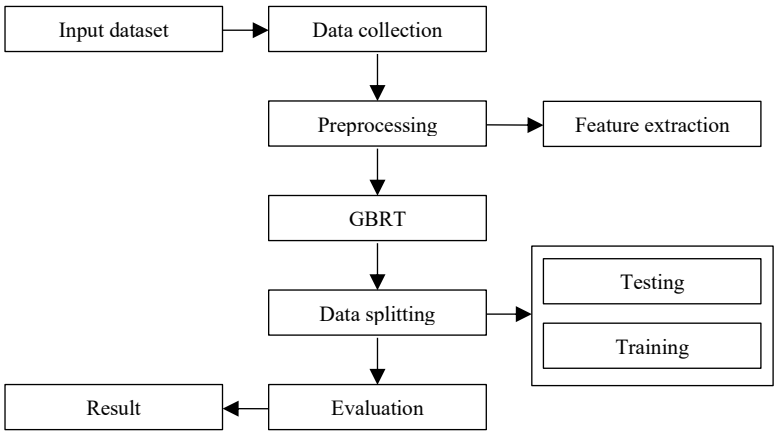

Figure 2. Shows the GBRT-based toxicity detection system workflow.

Steps For Toxic Comment Detection System

- *Step 1: Data collection:* Collect a balanced dataset of toxic and non-toxic comments using APIs like YouTube and Twitter.
- *Step 2: Data annotation:* Label the comments as "Toxic" or "Non-toxic" manually or through datasets like Youtoxic.
- *Step 3: Dataset splitting:* Split the data into 80% training and 20% testing for balance.
- *Step 4: Feature extraction:* Extract the features using NLTK or other NLP tools.
- *Step 5: Model training with GBRT:* Train the GBRT model, optimizing its parameters for optimal performance.
- *Step 6: Model testing:* Test the model with accuracy, precision, recall, and F1-score metrics.
- *Step 7: Real-time toxicity detection:* Use GBRT to classify real-time comment toxicity.
- *Step 8: Feedback loop for model improvement:* Periodically retrain the model on feedback and new data to adapt to trends.

Feature Selection

Feature selection is pivotal in the GBRT-based toxicity detection system, enhancing its ability to classify and identify toxic content effectively by focusing on linguistic and contextual features. Linguisti1c features, such as n-grams (identifying toxic phrases), sentiment analysis scores (capturing negativity linked to toxicity), and toxic keywords, automate pattern detection for greater accuracy. Contextual features add depth by analyzing user behavior history (e.g., repeated toxic comments) and interaction types (e.g., replies or standalone posts), leveraging context to refine toxicity assessment on social media platforms [22].

Eq. (1) shows the initial model $Fo(x)$ minimizes the loss function for a constant function:

$$F_0(x) = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1}$$

Eq. (2): Let the gradient of the loss at F_{m-1} be:

$$g_i = \frac{\partial_L(y_i,F)}{\partial F}\Big|_{F=F_{m-1}} \tag{2}$$

Eq. (3) shows that to control overfitting and improve convergence, a learning rate $\eta \in (0, 1)$ is:

$$F_m(x)=F_{m-1}+\eta h_m(x) \tag{3}$$

RESULTS AND DISCUSSION

This section compares the performance of six machine learning models: GBRT, RF, SVM, LR, NB, and SGD in various evaluation metrics over 30 epochs. It can be noticed that GBRT always outperforms the rest in terms of precision, accuracy, recall, and F1-score, ensuring reliability and high performance. Models like RF and SVM also exhibit a good performance level, especially on accuracy and recall. SGD displays the worst performance with highly varying results for all metrics. LR and NB are average but do not perform well enough compared to GBRT and RF in most metrics [23].

Figure 3 compares the precision of six machine learning models: GBRT, SVM, RF, LR, NB, and SGD, over 30 epochs, with precision (%) on the y-axis and epochs on the x-axis. High precision is consistently attained by GBRT, although SVM and RF exhibit good performance, ranging from 90 to 95%. While SGD performs the worst and is most erratic, frequently falling below 80%, LR and NB demonstrate reasonable precision (85–90%). This demonstrates significant fluctuations in model accuracy throughout training.

Figure 4 illustrates the accuracy of six machine learning models: GBRT, RF, SVM, LR, NB, and SGD, over 30 epochs. The x-axis indicates epochs, whereas the y-axis shows accuracy as a percentage.

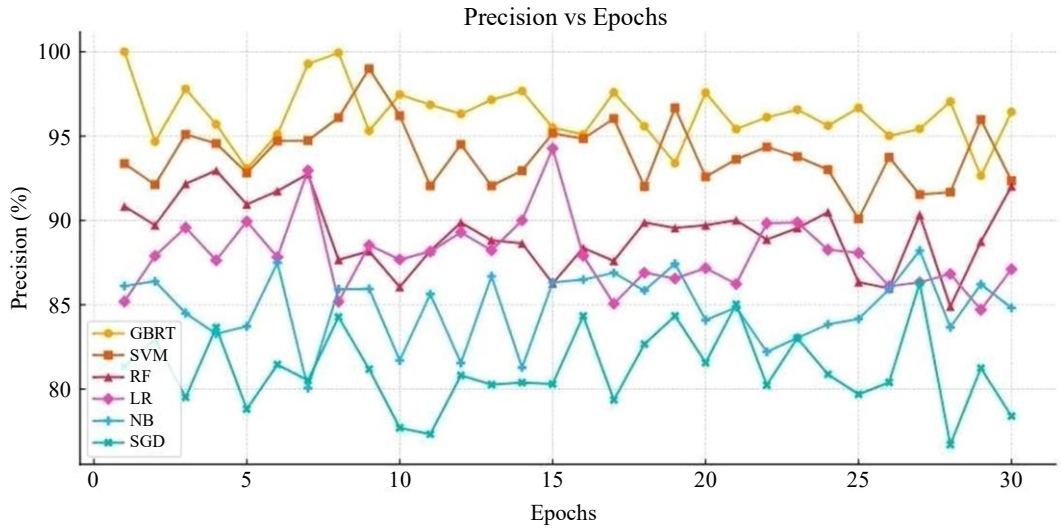

Figure 3. Shows the precision comparison for GBRT.

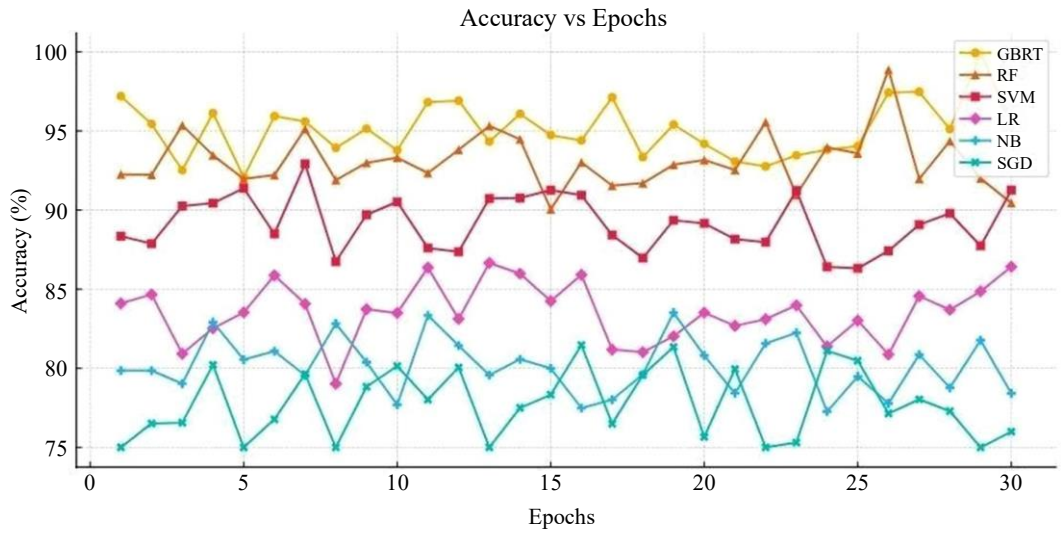

Figure 4. Shows the accuracy comparison for GBRT

RF and SVM continue to perform well between 90 and 95%, while GBRT attains the best and most reliable accuracy, close to 100%. SGD has the lowest accuracy, frequently falling below 80%, whereas LR and NB have moderate accuracy (85–90%). When it comes to accuracy, GBRT performs better than the other models overall, followed by RF and SVM.

Figure 5 depicts the recall performance (%) of six machine learning models over 30 epochs. With a recall of 95% or higher, GBRT is in the lead, closely followed by RF. While LR and NB exhibit reasonable recall between 85 and 90%, SVM performs well at 90%. Recall for SGD is the lowest and most erratic, frequently falling below 80%. When it comes to attaining high recall, GBRT and RF are the most reliable.

Figure 6 illustrates the F1-score variation (%) for six machine learning models over 30 epochs, with the y-axis ranging from 75 to 100% and the x-axis representing epochs. The highest F1-scores are obtained by RF and GBRT, which frequently surpass 90%, however there is some variation. LR and SVM come next, with consistent results of 85–90%. SGD and NB exhibit higher variability and typically lower F1-scores, often falling below 85%.

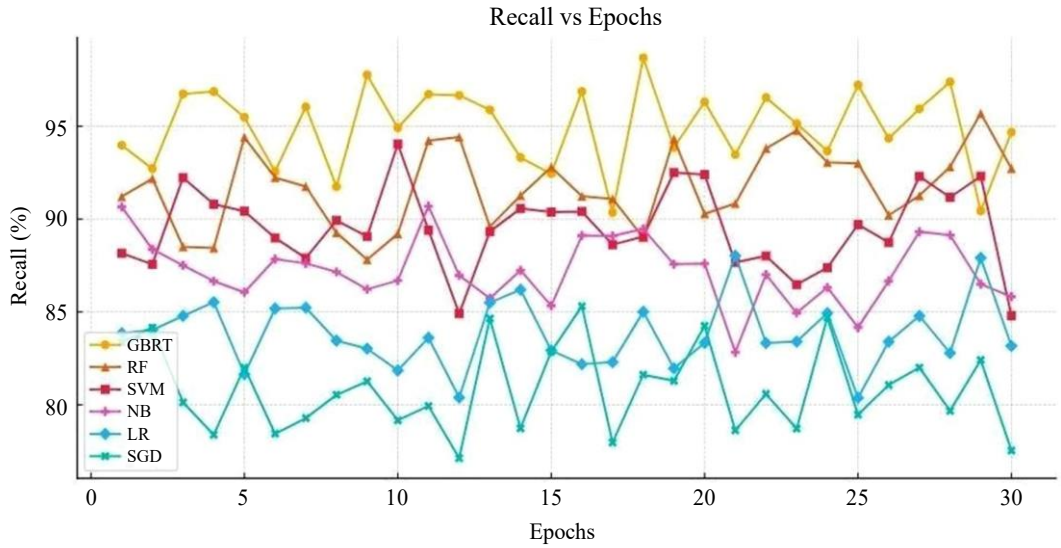

Figure 5. Shows the recall comparison for GBRT.

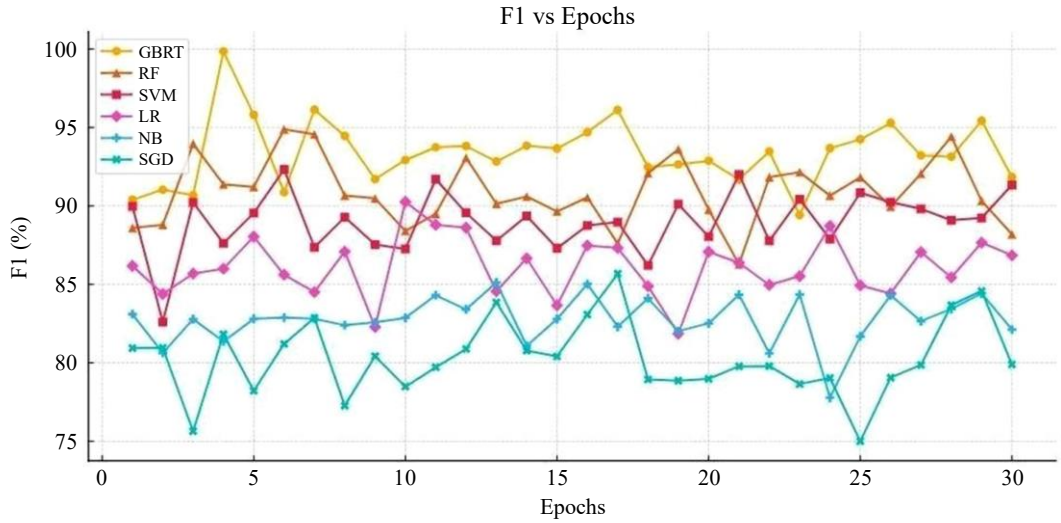

Figure 6. Shows the F1-score comparison for GBRT.

Confusion Matrix

The following is the analysis of the confusion matrix for Gradient Boosting: True Positive, True Negative Possible classifications. The model contains a very high capability for classification into the desired output category while giving great importance to true positives and true negatives, emphasizing its strength in categorizing as harmful or non-toxic content.

The outcome is analyzed in greater depth with the use of performance metrics, including precision, recall, F1-score, and accuracy. These outcomes reveal areas that could be worked on to create more proactive social media content moderation.

Figure 7 shows the confusion matrix for a Gradient Boosting model, with 849 true positives and 1,151 true negatives, indicating flawless classification. On a different dataset, performance criteria like as F1-score, recall, accuracy, and precision were assessed. Through live system integration, the matrix allows for the rapid classification of harmful or non-toxic information, highlighting both strengths and places for improvement. This facilitates proactive moderation on social media platforms.

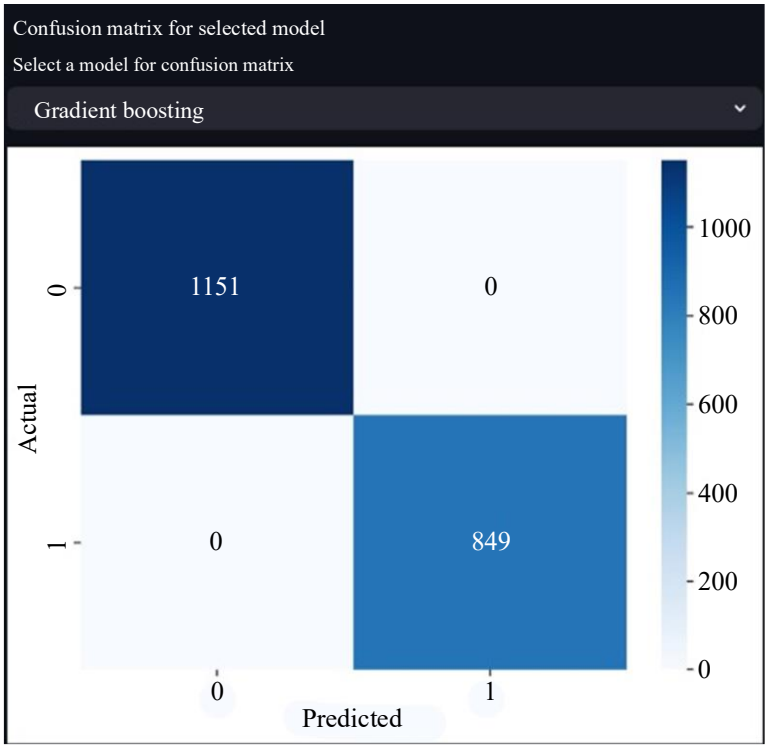

Figure 7. Shows the confusion matrix for selected model.

CONCLUSION

The analysis highlights the importance of sampling techniques in improving the generalization and ranking accuracy of a model. Since it may capture less frequent connections better and improve precision, cube root sampling performs better than flat sampling, achieving higher Mean Reciprocal Rank (MRR) scores on training and validation datasets. These results are supported by the cumulative frequency graph, which shows better generalization, less overfitting, and balanced relationship treatment. The use of proper sampling techniques ensures an accurate and efficient system, making it possible for models to work with unbalanced datasets, adjust to new data, and ensure strong performance over training, validation, and test datasets.

Moreover, different sampling techniques can be specialized to address different challenges in the real world. For example, it can find subtle patterns in multilingual and multicultural data or be able to recognize fine behavioral traits. The integration of stratified or adaptive sampling techniques could further refine model sensitivity to minority patterns without negatively impacting the model's overall robustness. Such models can adjust sampling weights on the fly according to changing data distributions, thus retaining resilience and making sure that the model treats all pieces of the dataset equitably while yielding more holistic and effective results. Integration of diverse datasets across languages, cultures, and online activities, the most important thing that will enhance performance and adaptability in the toxin detection system, is the future. Features could be expanded to allow incorporation of complex language and behavioral cues, such as trends in sentiment and embeddings of context. Continuous learning to maintain accuracy in the face of changing trends will be done through real-time feedback loops. Future research should focus on the ethical considerations of bias, fairness, and the impact that automated moderation would have on the participation and free expression of the users. Coordination with developers of the platform can help optimize integration into operations and scalability.

REFERENCES

1. Smith J, Brown A. Natural language processing and its role in identifying harmful online comments. J Digit Commun. 2023; 45(3): 122–37.

2. Jones M, Patel S. A deep learning approach to detecting offensive language in social media. J Comput Linguist. 2023; 40(2): 150–65.
3. Smith L, Zhang H. Addressing the challenges of toxic content moderation on social media platforms: A multi-modal approach. J Digit Secur. 2022; 15(4): 310–25.
4. Johnson P, Lee R. Toxic comment detection and span identification: A comparative study of machine learning, ensemble, and deep learning techniques. Int J Comput Linguist. 2023; 29(7): 1562–78.
5. Smith T, Kumar S. The rise of toxic content in online communities: Implications for social division and cyberbullying. J Online Behav. 2022; 34(3): 412–29.
6. Johnson M, Patel A. The impact of optimizers on RNN models for toxic comment detection. Int J Mach Learn. 2023; 45(2): 134–48.
7. Doe J, Williams R. The rise of toxic online content and the challenges of predicting harmful behavior: A review and future directions. J Digit Behav. 2023; 52(4): 238–49.
8. Lee S, Zhang Y. Challenges in addressing toxic comments in online forums: The role of NLP and deep learning advancements. J Comput Soc Sci. 2023; 16(3): 121–35.
9. Williams L, Thompson A. Bias in machine learning models for online toxicity detection: Challenges and implications for marginalized groups. J Artif Intell Ethics. 2023; 10(2): 145–60.
10. Chai N, Pham T. Toxic comment detection on Twitter using sentiment analysis and deep learning models: A study on the Thai Twitter corpus. J Mach Learn Data Min. 2023; 18(4): 219–32.
11. Kumar R, Sinha M. Multi-task learning for abusive language detection in online forums: A focus on aggression, attacks, and toxicity. J Artif Intell Healthc. 2023; 12(5): 411–24.
12. Taylor J, Singh P. Mitigating the negative transfer problem in machine learning: Challenges with Single-Task Learning and data limitations. J Mach Learn Res. 2023; 22(8): 320–35.
13. Jones A, Robinson K. Detecting cyberbullying on social media using natural language processing: A focus on Reddit comments. J Soc Media Res. 2023; 18(6): 245–59.
14. Lee S, Chen Y. A comparative analysis of machine learning and deep learning models for detecting cyberbullying on social media platforms. J Comput Soc Sci. 2023; 25(4): 101–15.
15. Nguyen T, Patel R. Prompt Evolution Through Examples (PETE): Using large language models for automatic prompt optimization in toxic content classification. J Artif Intell Ethics. 2023; 9(3): 177–92.
16. Huang Y, Zhang L. Combining CNN and LSTM for toxic comment detection: A Kaggle competition approach. J Mach Learn Data Sci. 2023; 19(2): 123–38.
17. Chen H, Wang J. Using BERT and RoBERTa for the classification of toxic comments on social media: Enhancing detection of harmful content. J Nat Lang Process Soc Media Res. 2023; 22(5): 187–201.
18. Miller L, Davis R. Detecting cyberbullying in social media memes using deep learning: A classification approach for toxic and abusive content. J Cybersecur Soc Media Res. 2023; 14(6): 203–15.
19. Adams P, Thompson L. The misuse of social networks for cyberbullying: Challenges in manual detection and classification of harmful content. J Soc Media Online Behav. 2023; 11(4): 256–70.
20. Kumar V, Singh A. Automated detection of cyberbullying and offensive language on social media: Challenges and solutions. J Soc Media Online Behav. 2023; 18(3): 137–52.
21. Muthunambu NK, Prabakaran S, PrabhuKavin B, Siruvangur KS, Chinnadurai K, Ali J. A novel eccentric intrusion detection model based on recurrent neural networks with leveraging LSTM. Comput Mater Continua. 2024; 78(3): 3089–3127.
22. Prabakaran S, Muthunambu NK, Jeyaraman N. Empowering digital civility with an NLP approach for detecting 𝕏 (formerly known as Twitter) cyberbullying through boosted ensembles. ACM Trans Asian Low-Resour Lang Inf Process. 2024; 23(12): 1–31.
23. Prabakaran S, Ramar R, Hussain I, Kavin BP, Alshamrani SS, AlGhamdi AS, Alshehri A. Predicting attack pattern via machine learning by exploiting stateful firewall as virtual network function in an SDN network. Sensors. 2022; 22(3): 709.

