A Comparative Analysis of Classical and Deep Learning Models for Toxic Comment Classification

Duc Nguyen Huu Dang,
CYBERSOFT, 15 Tran Khac Chan, Tan Dinh, Ho Chi Minh, Vietnam
FPT University, Ho Chi Minh Campus, Vietnam
huuduc01042006@gmail.com

Abstract

Toxic comments on online platforms present significant challenges to maintaining healthy online communities, making automatic toxic comment classification an important task in natural language processing. In this report, we focus on the application of a Multi-layer Perceptron (MLP) model for toxic comment classification and compare its performance with other machine learning approaches. Text features are extracted using standard vectorization techniques, and the MLP model is trained and evaluated on a labeled dataset of online comments. Performance is assessed using common evaluation metrics, including accuracy, precision, recall, and F1-score. Experimental results indicate that the MLP model achieves competitive performance compared to other models, demonstrating its effectiveness for toxic comment classification while maintaining a relatively simple architecture.

Index Terms

Online comments classification, Machine Learning, Multi-layer Perceptron.

I. Introduction

The rapid growth of online social platforms has led to a significant increase in user-generated content, including comments and discussions. Along with these developments, toxic comments such as hate speech, insults, and abusive language have become more prevalent, negatively affecting online communities and user experience. Automatic toxic comment classification has therefore emerged as an important task in natural language processing (NLP).

Early approaches to toxic comment detection mainly relied on traditional machine learning models combined with text feature extraction techniques such as bag-of-words and TF-IDF. More recently, deep learning–based models have demonstrated strong performance in text classification tasks. However, these models often involve complex architectures and high computational costs, which may not be suitable for small-scale systems or limited-resource environments.

In this report, we focus on the Multi-layer Perceptron (MLP) model for toxic comment classification and evaluate its effectiveness through a comparative study with other machine learning classifiers. Using vectorized text representations, the models are trained and evaluated on a labeled dataset of online comments. Performance is measured using standard metrics, including accuracy, precision, recall, and F1-score. The objective of this work is to examine whether a relatively simple MLP architecture can achieve competitive performance while maintaining low model complexity.

The main contributions of this report can be summarized as follows:

1) An implementation of an MLP-based model for toxic comment classification.
2) A comparative evaluation between MLP and other machine learning approaches under the same experimental settings.
3) An empirical analysis of the strengths and limitations of MLP for toxic comment detection.

II. METHOD

A. Dataset and Preprocessing

1) Dataset: Online comments labelled with six toxicity categories—toxic, severe-toxic, obscene, threat, insult, and identity-hate—make up the Jigsaw Toxic Comment dataset. Each comment can be classified into numerous categories at the same time, resulting in a multi-label classification problem. An initial analysis of the Fig.1 and

Fig.2 reveals a significant class imbalance across the six toxicity labels. The majority of comments are non-toxic, while certain categories such as threat and severe-toxic appear far less frequently than others. This imbalance is a well-known characteristic of the Jigsaw dataset and poses challenges for model evaluation, as standard metrics such as accuracy may be biased toward majority classes. I further analyze the distribution of comment lengths in the dataset. Most comments are relatively short, with a small number of longer comments containing more complex language patterns. This observation supports the use of bag-of-words–based representations such as TF-IDF, which are effective for capturing keyword-level information in short text sequences.

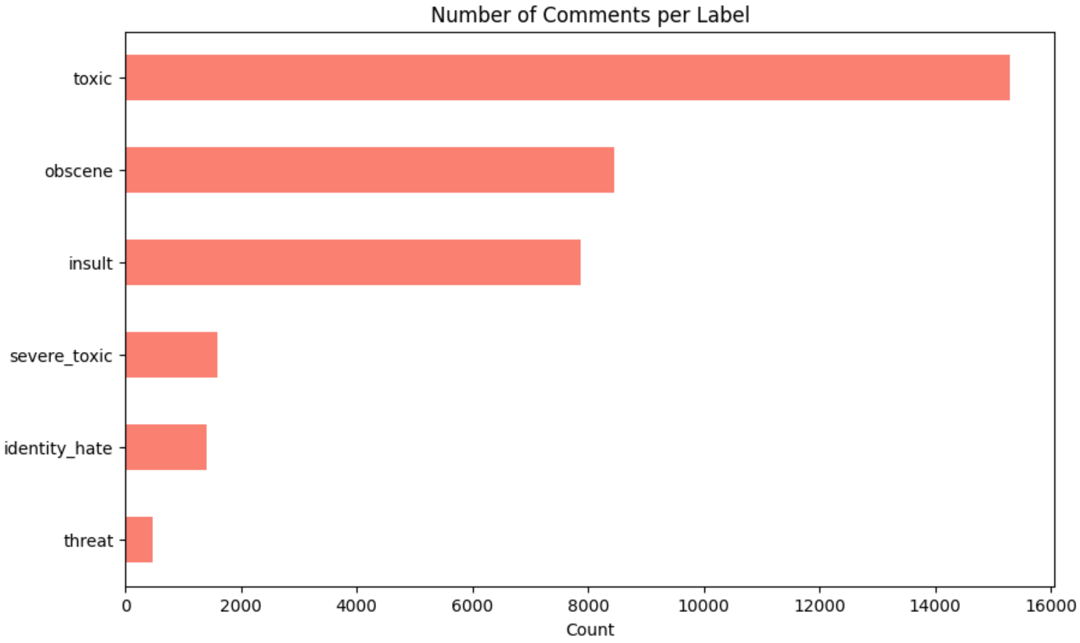

Fig. 1. Enter Caption

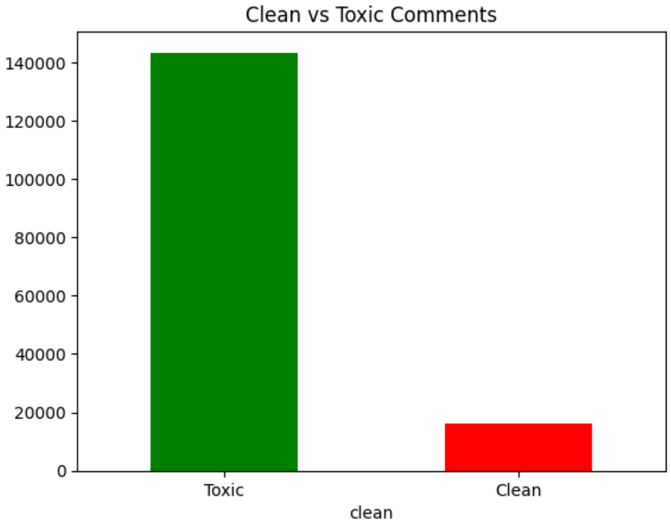

Fig. 2. Enter Caption

2) Preprocessing: Prior to model training, several preprocessing steps are applied to the raw text data in order to improve data quality and reduce noise. First, the id column is removed since it does not contain any semantic information relevant to the classification task. Missing values in the comment text are handled by replacing NaN entries with empty strings to ensure consistency during text processing. Next, the text is cleaned by converting all characters to lowercase and removing punctuation, special symbols, and unnecessary whitespace. Stopwords are then eliminated to reduce the influence of common but uninformative words. This step helps emphasize discriminative terms that are more relevant to toxic language detection. Finally, stemming is applied to normalize words by reducing them to their root forms. This process helps mitigate variations caused by different word inflections and contributes to a more compact and consistent feature representation. The resulting cleaned and normalized text is subsequently used for feature extraction and model training.

B. Baseline Models

1) Logistic Regression: Logistic Regression is employed as a linear baseline model for toxic comment classification. Text data are transformed into numerical representations using TF-IDF vectorization. The model is trained with L2 regularization to prevent overfitting and optimized using a gradient-based solver. Logistic Regression serves as a strong baseline due to its simplicity, efficiency, and effectiveness in high-dimensional sparse text representations.

2) Support Vector Machine: Support Vector Machine (SVM) is used as a margin-based classifier for comparison. Given the high-dimensional nature of text features, a linear kernel is adopted for computational efficiency. The SVM model aims to find an optimal separating hyperplane that maximizes the margin between toxic and non-toxic samples. This model is commonly used in text classification tasks and provides a robust baseline for performance comparison.

3) Naive Bayes: Naive Bayes is included as a probabilistic baseline model based on Bayes' theorem with a conditional independence assumption between features. Despite its simplicity, Naive Bayes has been shown to perform reasonably well on text classification tasks, particularly with TF-IDF features. Its fast training and inference make it a suitable baseline for large-scale text data.

4) Random Forest: Random Forest is an ensemble-based machine learning model that constructs multiple decision trees during training and aggregates their predictions. By combining multiple trees, Random Forest reduces overfitting and improves generalization compared to single decision trees. In this study, Random Forest is used to evaluate the effectiveness of ensemble learning methods for toxic comment classification.

C. Proposed Multi-layer Perceptron

The Multi-layer Perceptron (MLP) is employed as the main classification model in this study. The model takes TF-IDF feature vectors as input and is designed to address a multi-label toxic comment classification task involving six independent toxicity categories. The architecture of the MLP consists of an input layer followed by a single hidden fully connected layer with 512 neurons. Batch normalization is applied after the linear transformation to stabilize the feature distribution and improve training robustness. A ReLU activation function is used to introduce non-linearity, allowing the model to capture complex relationships within the high-dimensional textual features. To further reduce overfitting, dropout is incorporated with a probability of 0.3. The output layer contains six neurons, each corresponding to one toxicity label. Instead of using a softmax function, the model outputs independent logits for each class, enabling simultaneous prediction of multiple toxicity categories. During training, sigmoid activation is applied to transform logits into class-wise probabilities. Overall, this MLP architecture provides a balance between model simplicity and expressive capacity, making it suitable for learning discriminative patterns from sparse TF-IDF representations.

III. Results and Discussion

A. Model Settings

The training hyperparameters used for the proposed MLP model are summarized in Table X. The AdamW optimizer is adopted to update network parameters, as it effectively combines adaptive learning rates with decoupled weight decay, thereby improving generalization performance. The learning rate is set to 1×10^{-4} to ensure stable and smooth convergence during training. The loss function employed is BCEWithLogitsLoss, which is particularly suitable for multi-label classification tasks by integrating sigmoid activation with binary cross-entropy loss in a numerically stable manner. To alleviate overfitting, a dropout rate of 0.3 is applied to the hidden layer, randomly deactivating neurons during training. The ReLU activation function is used due to its computational efficiency and effectiveness in mitigating the vanishing gradient problem.(Table I)

TABLE I
TRAINING HYPERPARAMETERS OF MLP

HyperParameter	**Value**
Optimizer	AdamW
Learning rate	1×10^{-4}
Loss function	BCEWithLogitsLoss
Dropout	0.3
Activation	ReLU

B. *Performance evaluations*

Table II and Table III present the classification performance of different baseline models and the proposed MLP model on the Jigsaw toxic comment dataset, evaluated using Accuracy and F1-score, respectively, across six toxicity categories and overall performance. As shown in Table II, the proposed MLP model consistently achieves the highest accuracy across all toxicity classes. In particular, it attains an overall accuracy of **0.9902**, outperforming traditional machine learning models such as Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest. The improvement is especially notable in the *Toxic*, *Severe Toxic*, and *Identity Hate* categories, indicating that the MLP is more effective in capturing complex feature interactions from textual representations. Table III reports the mean F1-scores for all models. While baseline methods achieve reasonable accuracy, their F1-scores remain relatively low for minority classes such as *Threat* and *Severe Toxic*, reflecting challenges caused by class imbalance in the dataset. In contrast, the proposed MLP model demonstrates superior F1-score performance across all categories, achieving an overall F1-score of **0.80**. Notably, it achieves substantial gains in the *Threat* and *Identity Hate* classes, suggesting improved balance between precision and recall.

Overall, the experimental results indicate that the MLP model provides a strong trade-off between simplicity and effectiveness. Despite its relatively shallow architecture, the model significantly outperforms classical baselines, highlighting the benefit of non-linear feature learning for toxic comment classification.

TABLE II
ACCURACY (MEAN) OF TOXIC COMMENT CLASSIFICATION ON THE JIGSAW DATASET

Model	**Toxic**	**Sev. Toxic**	**Obscene**	**Threat**	**Insult**	**Id. Hate**	**Overall**
Logistic Regression	0.96	0.99	0.98	0.99	0.97	0.99	0.9817
Support Vector Machine	0.96	0.99	0.98	0.99	0.97	0.99	0.9817
Naive Bayes	0.95	0.99	0.97	0.99	0.97	0.99	0.9783
Random Forest	0.95	0.99	0.97	0.99	0.96	0.99	0.9766
MLP (Ours)	**0.977935**	**0.993739**	**0.989610**	**0.998565**	**0.985029**	**0.996497**	**0.9902**

TABLE III
F1-SCORE (MEAN) OF TOXIC COMMENT CLASSIFICATION ON THE JIGSAW DATASET

Model	**Toxic**	**Sev. Toxic**	**Obscene**	**Threat**	**Insult**	**Id. Hate**	**Overall**
Logistic Regression	0.86	0.68	0.87	0.59	0.81	0.64	0.74
Support Vector Machine	0.87	0.57	0.88	0.56	0.82	0.62	0.72
Naive Bayes	0.82	0.59	0.82	0.50	0.77	0.52	0.67
Random Forest	0.84	0.64	0.85	0.57	0.79	0.58	0.71
MLP (Ours)	**0.88**	0.68	**0.90**	**0.73**	**0.85**	**0.78**	**0.80**

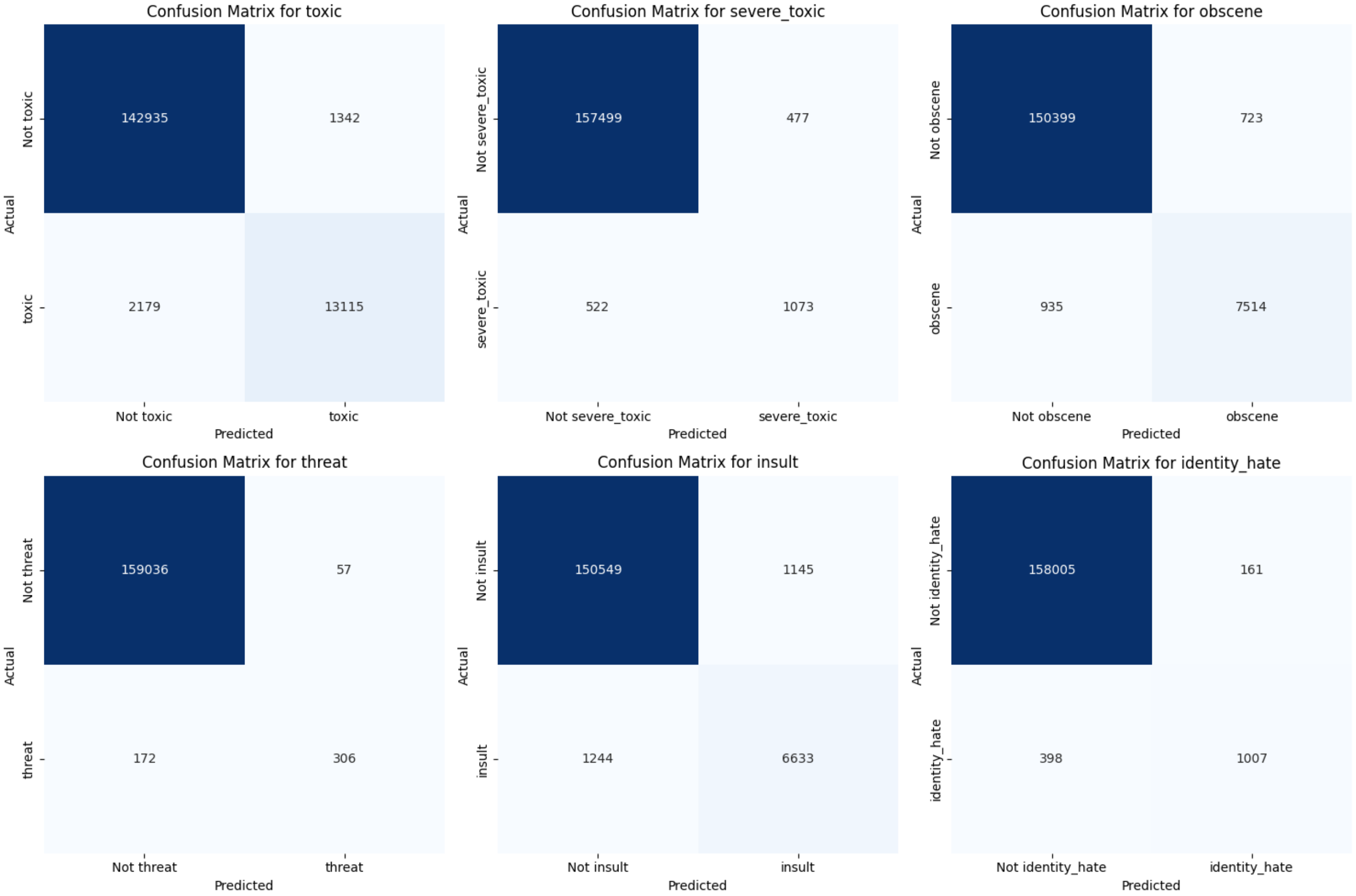

Fig. 3. Confusion Matrix

Figure 3 illustrates the confusion matrices of the proposed MLP model for each toxicity category on the Jigsaw dataset. Overall, the model demonstrates strong discriminative capability, with a large proportion of true negatives correctly classified across all classes, reflecting the inherent class imbalance of the dataset.For majority classes such as *toxic*, *obscene*, and *insult*, the model achieves high true positive rates with relatively low false negatives. In contrast, minority classes including *severe toxic*, *threat*, and *identity hate* exhibit higher misclassification rates, which is consistent with their limited number of positive samples. Nevertheless, the model is still able to correctly identify a substantial portion of these rare toxic instances, indicating robust performance in challenging low-frequency categories.

C. Discussion

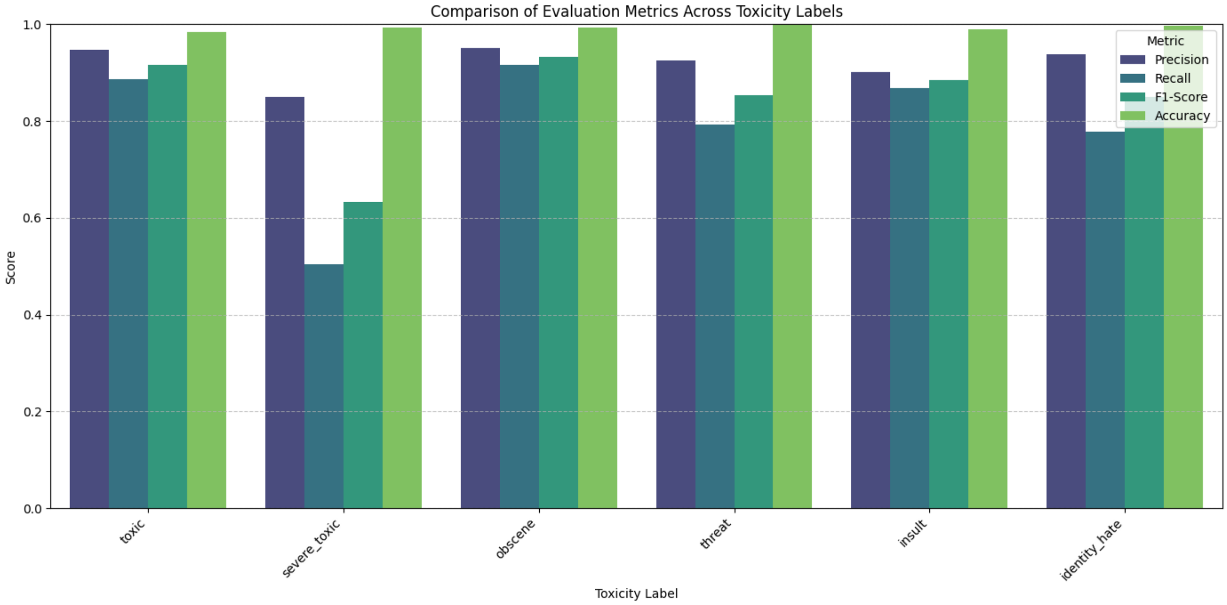

Fig. 4. Comparison of Precision, Recall, F1-score, and Accuracy across toxicity labels on the Jigsaw dataset.

Figure 4 presents a comparative visualization of Precision, Recall, F1-score, and Accuracy across different toxicity labels. The results show that majority classes such as *toxic* and *obscene* achieve consistently high performance, while minority classes, particularly *severe toxic* and *threat*, exhibit lower recall and F1-score, reflecting the impact of class imbalance.

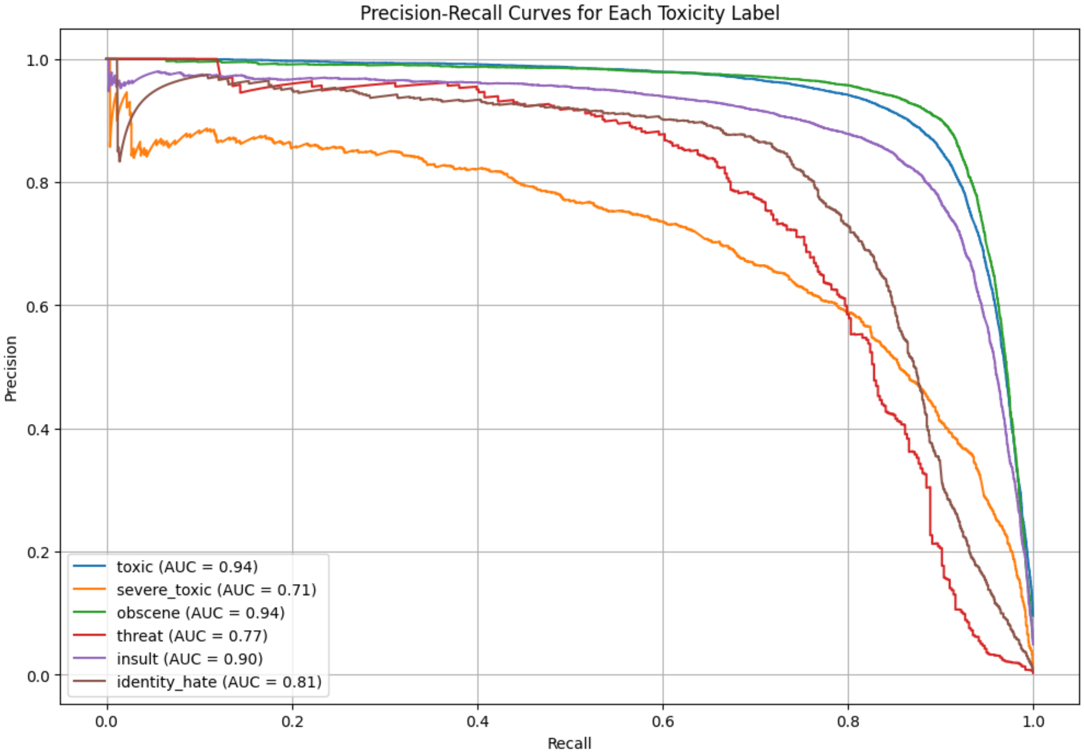

Fig. 5. PR Curves for Each Label

Figure 5 illustrates the Precision–Recall curves for each toxicity category. The proposed model maintains high precision over a wide recall range for frequent classes, whereas performance degrades more rapidly for rare labels. This behavior is consistent with the lower F1-scores observed in Table III.

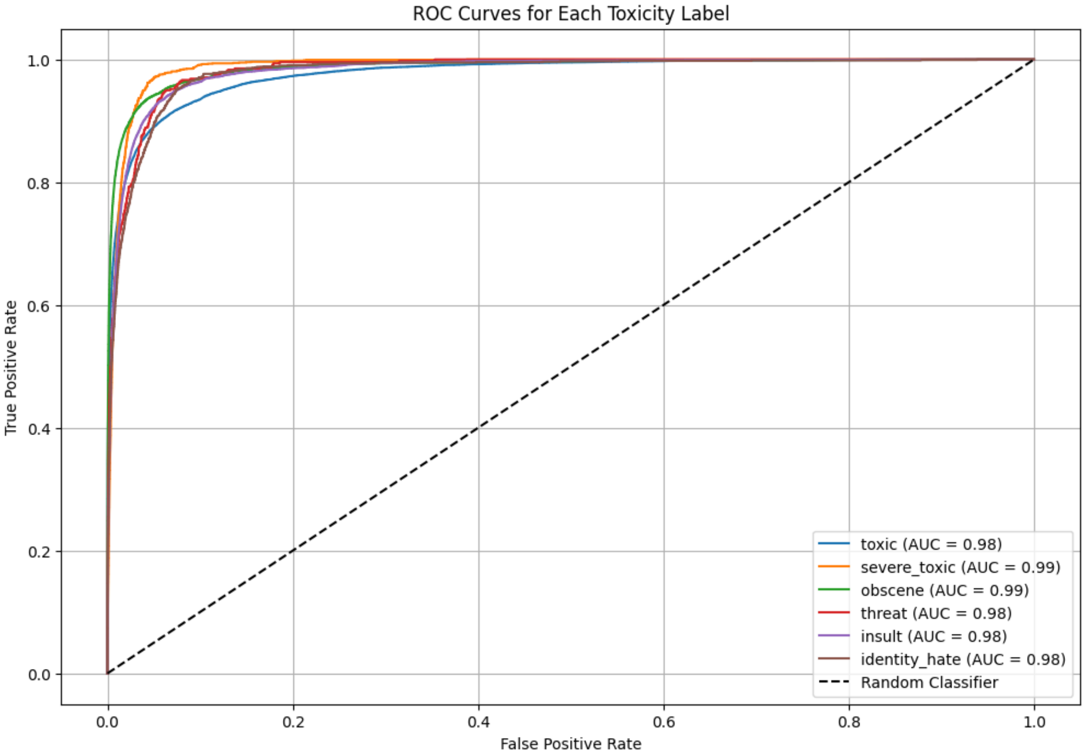

Fig. 6. ROC Curve for each Label

Figure 6 shows the ROC curves of all toxicity labels, where the model achieves high AUC values across categories, indicating strong overall discriminative capability.

IV. Conclusion

In this report, I investigated the toxic comment classification task on the Jigsaw dataset by comparing several traditional machine learning baselines with a Multi-layer Perceptron (MLP) model. Experimental results demonstrate that the proposed MLP consistently outperforms classical approaches such as Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest across all toxicity categories in terms of both accuracy and F1-score.Despite its relatively simple architecture, the MLP model effectively captures non-linear relationships in textual features, leading to improved performance, particularly on minority classes. Additional analyses using confusion matrices, Precision–Recall curves, and ROC curves further highlight the robustness of the proposed approach under severe class imbalance conditions. Overall, the results suggest that a lightweight neural network, when combined with proper preprocessing and training strategies, can serve as a strong and efficient solution for toxic comment classification. Future work may explore more advanced architectures and imbalance-aware training techniques to further enhance performance on rare toxicity labels.