International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 37-45
Received Date: 10/31/2025
Revised Date: 5/22/2026
Accepted Date: 5/29/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95699

Multi-Label Toxic Comment Detection Using BERT-Based NLP and Machine Learning

Binita Adhikari[1], Pratistha Sapkota[2], Rajad Shakya[3]

[1]*Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, Nepal, binitaa.2003@gmail.com*
[2]*Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, Nepal, pratistha.sapkota123@gmail.com*
[3]*Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, Nepal, shakyarajad1@gmail.com*

Abstract

The rapid growth of online communication platforms has significantly increased the prevalence of toxic and abusive language, creating major challenges for content moderation and online safety. Traditional keyword-based and machine learning approaches often fail to capture contextual and semantic nuances present in toxic comments. This study proposes a BERT-based toxic comment detection system using Natural Language Processing (NLP) and deep learning techniques for multi-label classification of online comments. The proposed model classifies comments into six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate using the Jigsaw Toxic Comment Classification Challenge dataset. The methodology includes data preprocessing, tokenization using the BERT tokenizer, and fine-tuning of the pre-trained BERT-base-uncased model. Experimental evaluation demonstrates that the proposed model achieved a micro-average F1-score of 0.7354, outperforming traditional machine learning approaches such as Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machine (SVM). The results indicate that transformer-based architectures effectively capture contextual relationships and implicit toxic expressions compared to conventional methods. However, performance on minority classes such as threat and identity hate remained limited because of severe dataset imbalance. The findings demonstrate the effectiveness of BERT for context-aware toxic comment classification and highlight its potential application in automated moderation systems for safer online communication platforms.

Keywords: Natural Language Processing, Machine Learning, BERT, Toxic Comment Detection, Multi-Label Classification, Online Safety.

1. Introduction

With the rapid expansion of internet users and user-generated content, managing harmful and offensive language online has become increasingly challenging. Traditional keyword-based and conventional machine learning approaches are often insufficient for detecting context-dependent toxicity such as sarcasm, implicit abuse, and indirect hate speech. Manual moderation is not scalable, especially given the vast volume of comments across social media, forums, and other online platforms. The problem is made worse by the anonymity of the internet, which encourages people to use hurtful and abusive language. If left unchecked, this toxicity can impede constructive discourse, marginalize vulnerable groups, and create a hostile online environment.

Modern deep learning and natural language processing (NLP) based methods have gained popularity as a means to addressing these issues. Because of its capacity to comprehend context and semantic structure, one such model, BERT (Bidirectional Encoder Representations from Transformers), has demonstrated exceptional performance in a variety of text classification tasks. In this study, we present a BERT-based approach for toxic comment detection, which classifies comments into six categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The model is trained and evaluated using the Jigsaw Toxic Comment Classification Challenge dataset, a widely accepted benchmark for multi-label toxicity classification.

The motivating force of this study is the increasing toxicity on digital platforms. Traditional keyword-based methods are insufficient for capturing context-dependent toxicity like sarcasm or implicit abuse, traditional keyword-based or machine learning techniques are insufficient. BERT is a context-aware model that considers complete sentence structure, making it highly suitable for this task. In addition to improving user experience,

creating a precise and timely toxic comment detection system also helps create a more secure and civil online community.

In order to improve classification results, this study intends to preprocess and curate a high-quality dataset appropriate for training deep learning models like BERT, assess model performance using metrics like F1-score, recall, precision, and accuracy, and investigate optimization strategies like class balancing and threshold tuning. The ultimate objective is to create a trustworthy, up-to-date tool for identifying and eliminating harmful online language.

2. Related Works

Toxic comment detection and hate speech classification have gained significant research attention in recent years because of the increasing prevalence of harmful content on online platforms. Early studies primarily relied on traditional machine learning techniques combined with handcrafted textual features.

One of the pioneering studies in toxic language detection was conducted by Zeerak Waseem and Dirk Hovy, who introduced a Twitter dataset containing sexist, racist, and neutral tweets. (Waseem and Hovy, 2016) Their work utilized classical machine learning algorithms such as Logistic Regression and Naive Bayes along with character-level n-grams and linguistic features for hate speech detection. Although the study provided important insights into abusive language classification, the models struggled to capture deeper contextual relationships within text.

Subsequently, Pinkesh Badjatiya et al. proposed a deep learning-based approach using Long Short-Term Memory (LSTM) networks for hate speech detection on Twitter. The learned semantic embeddings were further combined with gradient boosting classifiers, resulting in improved classification performance compared to traditional feature-engineering approaches. This work demonstrated the effectiveness of sequence modeling techniques in understanding implicit toxic expressions. (Badjatiya et al., 2017)

Further improvements were introduced by Ziqi Zhang et al., who proposed a hybrid neural network architecture combining Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs). The CNN layers extracted local semantic patterns, while the GRU layers captured sequential dependencies within text. Their hybrid model achieved strong performance in detecting offensive and hateful language on social media platforms. (Zhang, Luo and Zhang, 2018)

A major advancement in toxic comment classification was introduced through the Jigsaw Toxic Comment Classification Challenge organized by Google Jigsaw in 2018. (Google Jigsaw, 2018) The challenge released a large-scale dataset containing Wikipedia comments annotated across six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The competition accelerated research in toxic language detection and encouraged the development of advanced transformer-based architectures.

Transformer-based models significantly improved toxic language classification performance. Jacob Devlin et al. (Devlin et al., 2019) introduced BERT, a bidirectional transformer model capable of learning deep contextual representations from text. BERT achieved state-of-the-art results across multiple NLP tasks because of its ability to analyze words using both left and right contextual information simultaneously. Complementing this, Mozafari, Farahbakhsh and Crespi (2019) applied BERT-based models to racial bias mitigation alongside hate speech detection in social media, demonstrating the model's adaptability across toxicity-related tasks.

Building upon transformer architectures, Zhilin Yang et al. proposed XLNet, which addressed certain limitations of BERT using generalized autoregressive pretraining techniques. Similarly, (Yang et al., 2019) Yinhan Liu et al. introduced RoBERTa, an optimized transformer architecture that achieved superior benchmark performance through improved pretraining strategies. (Liu et al., 2019) Recent studies have further emphasized the importance of high-quality datasets and robust deep learning architectures for toxic comment classification. Julian Risch and Rico Krestel demonstrated the effectiveness of transformer-based deep learning models for toxic comment detection in online discussions. (Risch and Krestel, 2020) Additionally, Bertie Vidgen and Leon Derczynski highlighted that poor dataset construction and annotation quality can negatively affect abusive language classification systems (Vidgen and Derczynski, 2020).

3. Model Architecture and Theory

This project utilizes the pre-trained BERT-base model as the core architecture for toxic comment classification. BERT (Bidirectional Encoder Representations from Transformers) generates deep bidirectional contextual representations using transformer encoder layers and self-attention mechanisms. Its ability to capture semantic and contextual relationships within text makes it highly effective for toxic comment classification tasks.

1. **Pre-trained BERT base (uncased)**: A 12-layer transformer encoder with roughly 110 million parameters was pre-trained using next sentence prediction tasks and masked language modeling on a sizable English corpus. As a result, the model produces robust general-purpose language representations that are fine-tuned for the specific classification task.
2. **Dropout Layer**: A dropout layer is applied to the BERT pooled output in order to minimize overfitting when fine-tuning. Dropout rates are commonly set between 0.1 and 0.3 to prevent overfitting during training.
3. **Dense Output Layer**: The 768-dimensional pooled BERT output is mapped to six output neurons, each of which represents a toxicity category, via a fully connected linear layer. By using a sigmoid activation function to create multi-label probabilities, each comment can be simultaneously classified into several toxicity categories.

The overall architecture of the proposed BERT-based toxic comment classification system is illustrated in Figure 1.

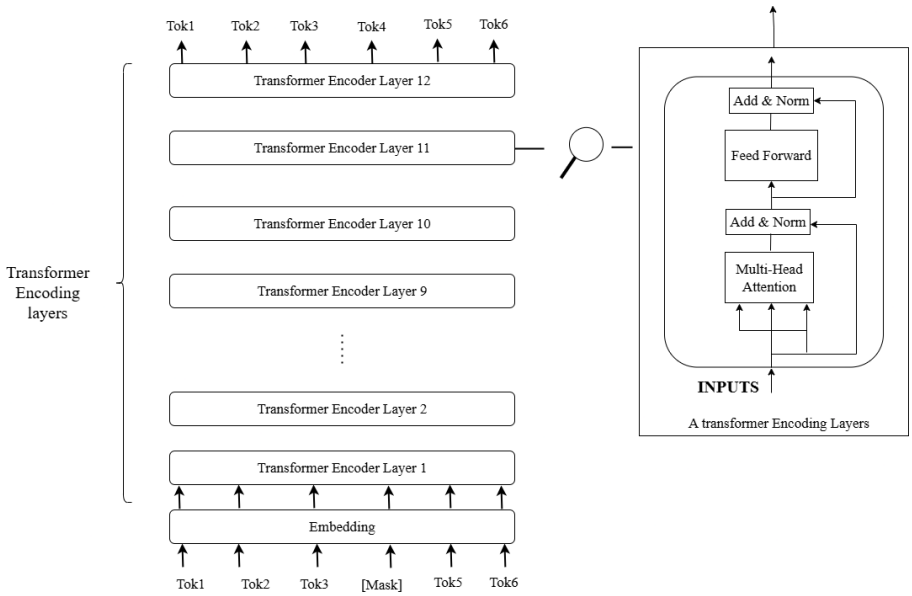

Figure 1: BERT base model architecture

3.1 *Mathematical Formulation in BERT*

The BERT model leverages the transformer encoder, which is fundamentally based on the self-attention mechanism.

- **Scaled Dot-Product Attention**:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k is the dimension of keys and Q, K, and V are the query, key, and value matrices, respectively.

- **Multi-Head Attention**:

$$\text{Multi Head}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\boldsymbol{head_1}, \ldots, \boldsymbol{head_h})\boldsymbol{W^O}$$

$$head_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where W^O is the output projection matrix and $\left(W_i^Q, W_i^K, W_i^V\right)$ are projection matrices for the i - th head.

- **Feed-Forward Network**:

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2$$

which is applied identically and independently to every position.

Formation of Query, Key, and Value matrices:

Given an input embedding matrix $\boldsymbol{X} \in \boldsymbol{R}^{n \times d_{model}}$, where n is the sequence length and $\boldsymbol{d_{model}}$

is the embedding dimension, the matrices Q, K, and V are obtained by linear projections:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V$$

Where W^Q, W^K, and $W^V \in R^{d_{model} \times d_k}$ are the learnable weight matrices. These projections transform the input embeddings into query, key, and value vectors, which are then used in the computation of attention methodology.

This study adopts a systematic multi-stage methodology for toxic comment classification using the BERT model. The overall process involves dataset collection and preparation, text preprocessing, tokenization, model fine-tuning, and performance evaluation. Each stage was carefully designed to improve the accuracy and robustness of the multi-label toxic comment detection system.

3.2 *Dataset Overview*

We used the Jigsaw Toxic Comment Classification Challenge dataset from Kaggle, which contains user comments from Wikipedia talk pages labeled for various toxicity categories. The dataset includes 159,571 comments in the training set and 63,978 in the test set. It is structured in CSV format and includes a unique identifier (ID), comment text, and binary labels for six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Each comment may belong to multiple classes, making this a multi-label classification problem. Preprocessing involved lowercasing, optional stopword removal and text cleaning (removing HTML tags, URLs, non-alphabetic characters).

Table 1: Label Distribution in the Jigsaw Dataset

Toxicity Type	Proportion in Dataset
Toxic	15.7%
Severe Toxic	1.0%
Obscene	10.2%
Threat	0.3%
Insult	5.5%
Identity Hate	1.4%

3.3 *Data Preprocessing*

Before training, we performed extensive preprocessing. Comments with missing or null text were removed to ensure data integrity. We combined the six binary target columns into a consolidated label column, which lists all the toxicity tags assigned to a given comment. The dataset was split into training and validation subsets using an 80-20 stratified split, preserving label distribution across both sets. This stratification is essential to prevent biased model evaluation, particularly because of class imbalance.

3.4 *Tokenization*

Tokenization was performed using the bert-base-uncased tokenizer of the HuggingFace 's Transformers library. The text of every comment was first lowercased for uniformity. The tokenizer subsequently split the text into subword tokens according to the WordPiece algorithm, as required by BERT. Each tokenized input was padded or truncated to a fixed sequence length of 256 tokens. The maximum token length of 256 was selected because most comments in the Jigsaw dataset fall within this range. This length provides a balance between preserving contextual information and maintaining computational efficiency during training. To help the model focus only on meaningful inputs during training, attention masks were created to distinguish valid tokens from padding.

3.5 *Custom Dataset & DataLoader*

We built a custom PyTorch Dataset class to handle tokenized inputs and corresponding labels in a memory-efficient way. This class returns input IDs, attention masks, and labels for each data point. We used the PyTorch DataLoader to enable mini-batch processing, data shuffling, and parallel loading, optimizing both memory usage and training time. A batch size of 16 was selected based on available GPU memory constraints and training stability. Smaller batch sizes help reduce memory consumption while maintaining effective gradient updates during fine-tuning.

3.6 *Training Configuration*

The model is trained using the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), which is suitable for multi label classification tasks. The AdamW optimizer, which has a linear learning rate of $\mathbf{2 \times 10^{-5}}$ was used, featuring a linear warm-up followed by gradual decay. The training process was run for three epochs. Each epoch included training on the entire dataset followed by evaluation on the validation set. These configurations were carefully selected to ensure stable fine-tuning of the BERT model.

3.7 *Evaluation Strategy*

To generate model predictions, sigmoid outputs were thresholded at a value of 0.5. Using both class-level and micro/macro averages, the model was evaluated with common classification metrics such as accuracy, precision, recall, and F1-score. Training and validation loss curves were analyzed to monitor convergence and identify potential overfitting during training.

Table 2: Model Parameters and Training Configuration

Parameter	Value
Base Model	BERT-base-uncased
Dropout Rate	0.3
Output Layer	Dense (6 neurons)
Loss Function	BCEWithLogitsLoss
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	3
Activation Function	Sigmoid
Evaluation Metrics	Accuracy, Precision, Recall, F1-score

4. Result and Discussion

4.1 Evaluation Metrics of BERT

The final performance of the BERT-based toxic comment classification model is summarized as follows:

- **Final Precision (Weighted):** 0.7189
- **Final Recall (Micro):** 0.7269
- **Final F1-Score (Micro):** 0.7354

Table 3: Evaluation Metrics for BERT Model

Label	Precision	Recall	F1-Score
Toxic	0.80	0.80	0.80
Severe Toxic	0.58	0.17	0.27
Obscene	0.75	0.81	0.78
Threat	0.00	0.00	0.00
Insult	0.65	0.78	0.71
Identity Hate	0.40	0.06	0.11
Micro Avg	0.74	0.73	0.74
Macro Avg	0.53	0.44	0.45
Weighted Avg	0.72	0.73	0.71
Samples Avg	0.07	0.07	0.06

These findings indicate that the proposed BERT model performed effectively for commonly occurring toxicity classes such as toxic, obscene, and insult. However, comparatively lower performance was observed for minority classes such as severe toxic, threat, and identity hate due to significant class imbalance in the dataset.

4.2 Comparison with Traditional Machine Learning Models

The proposed BERT model was compared with several traditional machine learning approaches. Support Vector Machine (SVM), Random Forest, Naive Bayes, and Logistic Regression were some of the models. The following table presents the average F1-scores across all labels for different models.

Table 4: Average F1-Scores of Traditional ML Models

Model	Average F1-Score
Logistic Regression	0.3864
Random Forest	0.3790
Naive Bayes	0.3085
SVM	**0.4651**

Among traditional approaches, SVM achieved the highest average F1-score (0.4651), demonstrating its relative strength in toxic text classification. However, its performance on less frequent classes like *severe toxic* and *identity hate* was notably weaker compared to BERT.

4.3 Overall Comparison

Compared to traditional machine learning models such as Logistic Regression and SVM, the proposed BERT-based model achieved significantly higher performance due to its bidirectional transformer architecture and contextual language understanding capability. Unlike traditional bag-of-words approaches, BERT captures semantic relationships between words as well as sentence-level contextual information, enabling more accurate detection of implicit toxicity and context-dependent abusive expressions.

The superior performance of BERT is mainly attributed to its ability to analyze words based on both their left and right context, which helps in understanding nuanced language patterns, sarcasm, and indirect hate speech more effectively than conventional machine learning techniques. In contrast, traditional models rely heavily on handcrafted features and shallow textual representations, limiting their capability to capture deeper semantic meaning.

However, the performance for minority classes such as severe toxic, threat, and identity hate remained comparatively low because of severe dataset imbalance. These categories contain substantially fewer training examples, limiting the model's ability to generalize effectively during evaluation. In particular, the threat category achieved an F1-score of 0 because the dataset contains only 0.3% threat-related samples. Due to this extreme imbalance, the model failed to learn meaningful threat-related linguistic patterns. Error analysis further revealed that short comments, sarcastic statements, and comments containing ambiguous or implicit hate speech frequently resulted in misclassification.

To further contextualize the effectiveness of the proposed model, its performance was compared with other state-of-the-art transformer-based toxic comment classification models reported in the literature. Although the proposed BERT model achieved a strong micro F1-score of 0.7354, more advanced transformer architectures such as RoBERTa and XLNet have reported higher benchmark performance on the Jigsaw dataset, indicating potential directions for future optimization and model enhancement.

Table 5: Comparison with State-of-the-Art Toxic Comment Classification Models

Model	Dataset	Performance Observation
SVM	Jigsaw Dataset	Lower performance than the proposed BERT model
Proposed BERT Model	Jigsaw Dataset	Achieved Micro F1-score of 0.7354
RoBERTa-based approaches (Liu et al., 2019)	Jigsaw Dataset	Reported higher benchmark performance
XLNet-based approaches (Yang et al., 2019)	Jigsaw Dataset	Reported higher benchmark performance

4.4 Training and Validation Loss Analysis

The training and validation loss curves shown in Figure 2 indicate stable model convergence during fine-tuning. The training loss decreased steadily across epochs, while the validation loss remained stable, suggesting effective learning with minimal overfitting.

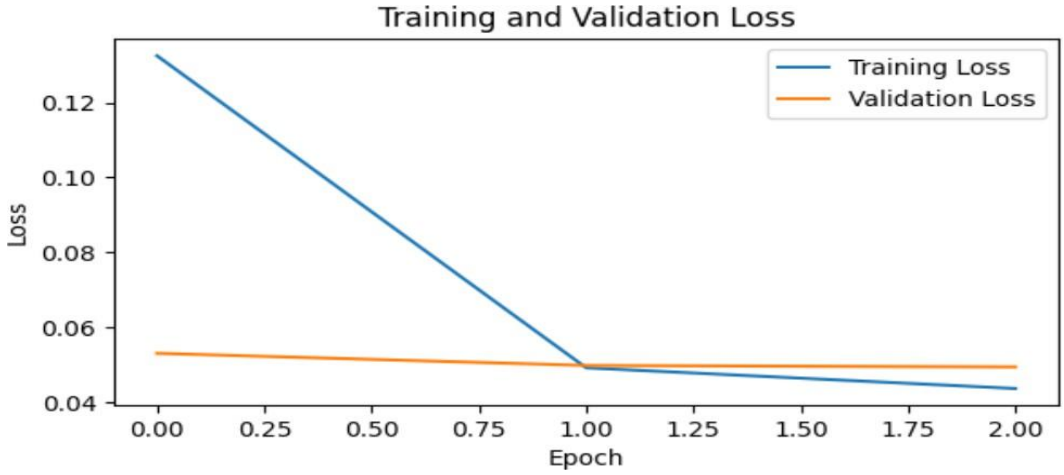

Figure 2: Training and Validation Loss Curves of the Proposed BERT Model

4.5 Confusion Matrix of BERT Model

The multi-label classification model's performance in each of the six categories is displayed in the confusion matrix. It shows how often predictions matched the true labels for each class, helping to identify both correct classifications and common misclassifications made by the model.

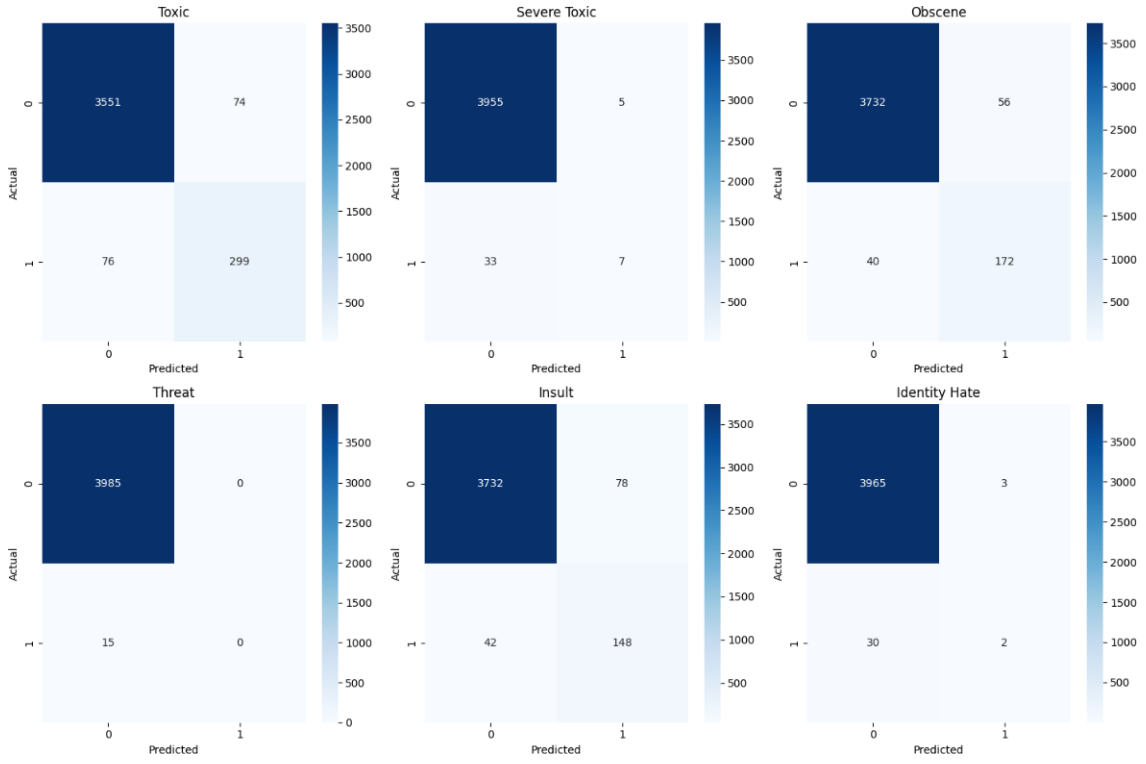

Figure 3: Confusion Matrix of the BERT-based Toxic Comment Classification Model

The toxic and obscene classes exhibit strong diagonal dominance, indicating accurate classification performance for majority categories with larger training sample sizes. In contrast, minority classes such as threat and identity hate show significantly weaker classification performance. Most threat-related comments

were incorrectly classified as non-toxic because the model was unable to learn sufficient threat related contextual features from the limited dataset samples.

4.6 Limitations

- The dataset is highly imbalanced, which reduced performance for minority classes such as *severe toxic*, *threat*, and *identity hate*.
- The proposed model is limited to English-language toxic comment detection and may not generalize well to multilingual datasets.
- BERT-based models require high computational resources and longer training time compared to traditional machine learning methods.
- The model sometimes struggles to detect sarcasm, implicit hate speech, and context-dependent toxicity.

5. Conclusion and Future Work

5.1 Conclusion

The project successfully classified multi-label toxic comments using BERT-based deep learning and attained strong performance in detecting varied types of toxic content such as hate speech, insults, and threats. The system's preprocessing, tokenization, and model fine-tuning through precise analysis significantly improved toxic comment classification performance online safety. The results validate that transformer-based NLP models are extremely effective in detecting the subtle and context-sensitive essence of online comments.

Overall, the study demonstrates the effectiveness of transformer-based NLP models for automated toxic language detection and highlights their potential application in building safer online communication environments.

5.2 Future Work

- Real-Time System Deployment – To implement real-time comment filtering, install the model in web platforms or live moderation software.
- Multilingual Extension – Scale the system to detect toxic content in Nepali or Hindi language using multilingual BERT.
- Bias Mitigation and Explainability – Overcome likely biases and incorporate explainable AI tools to improve transparency and fairness.
- Future work may incorporate focal loss, weighted loss functions, oversampling, and data augmentation techniques to address dataset imbalance. These methods can improve the detection performance of minority toxicity classes such as threat and identity hate.

References

Badjatiya, P., Gambäck, B., Gupta, M. and Varma, V. (2017) 'Deep learning for hate speech detection in tweets', *Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, Australia, pp. 759–760.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT*, pp. 4171–4186.

Google Jigsaw (2018) *Jigsaw toxic comment classification challenge* [online]. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge [Accessed: 1 June 2026].

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) 'RoBERTa: A robustly optimized BERT pretraining approach'. *arXiv preprint arXiv:1907.11692*.

Mozafari, M., Farahbakhsh, R. and Crespi, N. (2019) 'Hate speech detection and racial bias mitigation in social media based on BERT model'. In: *Complex Networks and Their Applications VIII*. Springer, pp. 405–417.

Risch, J. and Krestel, R. (2020) 'Toxic comment detection in online discussions using deep learning models', *Information Processing and Management*, 57(2), article 102099. https://doi.org/10.1016/j.ipm.2019.102099.

Vidgen, B. and Derczynski, L. (2020) 'Directions in abusive language training data: Garbage in, garbage out', *PLOS ONE*, 15(12), e0243300. https://doi.org/10.1371/journal.pone.0243300.

Waseem, Z. and Hovy, D. (2016) 'Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter', *Proceedings of the NAACL Student Research Workshop*, San Diego, CA, USA, pp. 88–93.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2019) 'XLNet: Generalized autoregressive pretraining for language understanding', *Advances in Neural Information Processing Systems*, 32.

Zhang, Z., Luo, J. and Zhang, J. (2018) 'Detecting hate speech on social media: A Convolution-GRU based deep neural network model', *IEEE Access*, 6, pp. 24552–24561.