Toxic Comment Classification Using Machine Learning and BERT

Name: Neha Kothavade
Course: Social Media Mining
IU Email: nkothav@iu.edu
Date: 27th Sept,2025

Abstract

Toxic online comments are a pressing challenge for digital platforms, harming conversations and necessitating automated moderation tools. This project develops and compares two machine learning approaches for toxicity classification: a Logistic Regression model with TF-IDF features and a fine-tuned BERT model. Using a dataset of 4,000 comments from Reddit, Twitter/X, and YouTube, I preprocess the data, implement both models, and evaluate their performance across accuracy, precision, recall, and F1. My findings show that the BERT model substantially outperforms Logistic Regression, particularly in recall and F1, demonstrating the importance of context-aware deep learning methods for toxicity detection.

1. Introduction And Background

Toxicity in online discussions can discourage participation, polarize communities, and spread hostility. Detecting such comments is essential for platforms aiming to maintain healthier online spaces. While classical machine learning methods like Logistic Regression provide interpretable baselines, they often fail to capture linguistic nuance such as sarcasm or multi-word expressions.

Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) represent a significant advancement in NLP by leveraging deep contextual embeddings. This report explores the effectiveness of traditional versus modern approaches by implementing and comparing:

1. Logistic Regression with TF-IDF features
2. Fine-tuned BERT model

The evaluation provides insights into which method better balances precision and recall in detecting toxic comments.

2. Dataset And Methods

To classify toxic comments, this section introduces the dataset of social media posts and the methods used to model their toxicity. I begin by outlining the key characteristics of the dataset and how the final labels were determined from human annotations. This provides the foundation for understanding how the training and evaluation process is structured.

I then describe the preprocessing steps that transform the raw text into a format suitable for machine learning. Finally, I present the two modelling strategies implemented in this project: a Logistic Regression model with TF-IDF features and a fine-tuned BERT model. Each approach is explained in terms of its design, implementation, and the unique strengths and challenges it brings to the task of toxicity detection.

2.1 Dataset

The dataset contains 4,000 comments from Reddit, Twitter/X, and YouTube. Each entry includes the following attributes:

- **text:** comment content
- **parent_comment:** original comment replied to (if any)
- **article_title & article_url:** contextual metadata
- **platform & platform_id:** source and identifier
- **composite_toxic:** five independent human annotations (toxic = true/false)

For supervised training, I derived ground truth labels using majority voting across the five annotations.

The dataset is split into:

- **Training set:** labeled comments with human-annotated toxicity scores.
- **Test set:** unlabeled comments requiring predictions.

2.2 Data Processing

Before training the models, several preprocessing steps were applied to transform the raw comments into a form suitable for machine learning.

The first step was **text cleaning**, where all characters were converted to lowercase to reduce vocabulary redundancy. Punctuation marks, digits, and extra white spaces were removed to minimize noise and ensure uniformity across comments from different platforms. This step helped standardize the text while preserving the semantic meaning of the comments.

Next, I determined the **ground truth labels** using **majority voting** from the five human annotations associated with each comment. A comment was marked as toxic if more than half of the annotators labeled it toxic. This approach reduced individual annotator bias and provided a consistent labeling scheme for supervised learning.

The dataset was then split into **training and validation sets** in an 80:20 ratio. The training set was used to fit the models, while the validation set allowed evaluation of model performance on unseen data, ensuring the models could generalize effectively.

Finally, I applied **feature extraction** for the Logistic Regression model using **TF-IDF vectorization**. Both unigrams and bigrams were considered, and the vocabulary size was capped at 5,000 features to balance expressiveness with computational efficiency. For the BERT model, raw text was tokenized using the pre-trained tokenizer associated with the model, preserving contextual information without manual feature engineering.

2.3 Machine Learning Methods

- ### Baseline TF IDF with Logistic Regression

Model Description:
This baseline model combines TF-IDF feature extraction with Logistic Regression. TF-IDF converts each comment into a weighted vector, where words that occur frequently within a comment but less often across the entire dataset receive higher weight. I used an n-gram range suitable for short text (unigrams and bigrams) and restricted the vocabulary size to 5,000 features to balance informativeness with efficiency. Logistic Regression modeled the probability of the toxic class with a sigmoid link, minimizing cross-entropy loss under L2 regularization.

Training Setup:
The model was trained with the scikit-learn implementation, using the default solver and regularization strength. I experimented with class weighting to address imbalance in toxic versus non-toxic comments. The prediction threshold was selected on the validation split by sweeping for the value that maximized F1 score rather than fixing it at 0.5, since maximizing F1 better balances precision and recall.

Strengths:
This approach is computationally efficient and easy to interpret, as feature weights directly indicate the importance of specific words or phrases. It works particularly well for short and explicit toxic expressions where certain terms act as strong indicators of toxicity.

Limitations:
The model cannot capture deeper semantics such as sarcasm, subtle toxicity, or the contextual interplay of words. It also struggles with longer sentences where toxicity depends on phrase-level meaning rather than individual terms, leading to lower recall.

- **Fine-Tuned BERT model**

Model Description:
The second model is based on BERT (Bidirectional Encoder Representations from Transformers), a transformer-based deep learning model that captures context and long-range dependencies in text. I used the pre-trained bert-base-uncased model from HuggingFace and added a linear classification head on top of the [CLS] token representation to predict toxicity.

Training Setup:
Comments were tokenized with WordPiece, truncated to a fixed maximum sequence length, and padded dynamically with attention masks. Training was performed for three epochs with a batch size of 8, using the AdamW optimizer and a learning rate of 2e-5. Mixed-precision training and gradient accumulation were applied for efficiency. Evaluation occurred at the end of each epoch, and probabilities for the toxic class were taken from the softmax over the two logits. As with Logistic Regression, I tuned the decision threshold on the validation set to maximize F1 score.

Strengths:
BERT captures contextual meaning within sentences, enabling it to detect nuanced or indirect toxicity that traditional models miss. It can represent long-range dependencies, handle multi-token expressions, and leverage pretrained knowledge to generalize effectively even with a relatively small dataset.

Limitations:
The model is computationally expensive, requiring GPUs for efficient training and inference. It is also more sensitive to hyperparameter settings, and can occasionally produce false positives by overinterpreting context. These factors make BERT more resource-intensive and harder to tune compared to simpler baselines.

3. Evaluations And Findings

I evaluated both models using four common metrics for classification tasks. Accuracy measures the overall correctness of predictions across all comments. Precision indicates how many of the comments predicted as toxic were actually toxic. Recall measures how many of the truly toxic comments were successfully detected. F1-score, the harmonic mean of precision and recall, balances these two aspects to give a single measure of effectiveness.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	77.0%	55.1%	55.6%	55.3%
BERT Model	80.4%	61.7%	62.0%	61.8%

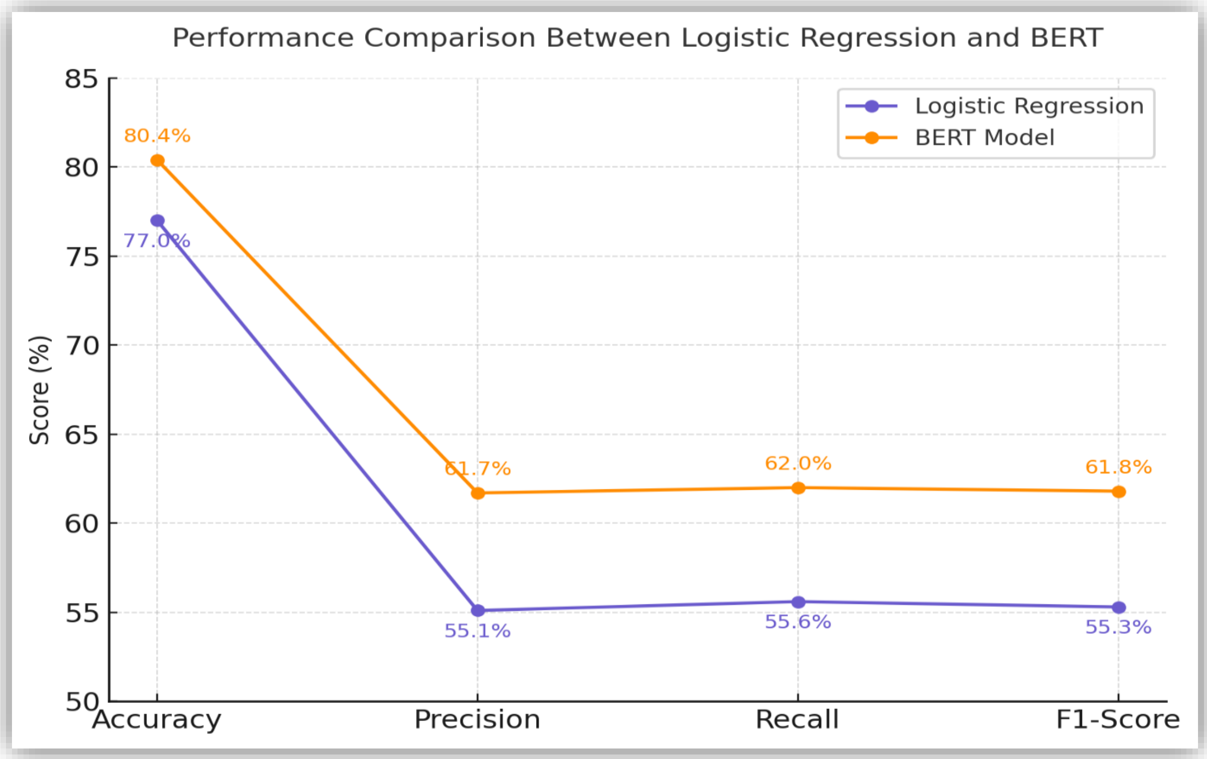

Figure 1. Line chart comparing Logistic Regression and fine-tuned BERT model performance across Accuracy, Precision, Recall, and F1-Score. The BERT model consistently achieves higher values across all metrics.

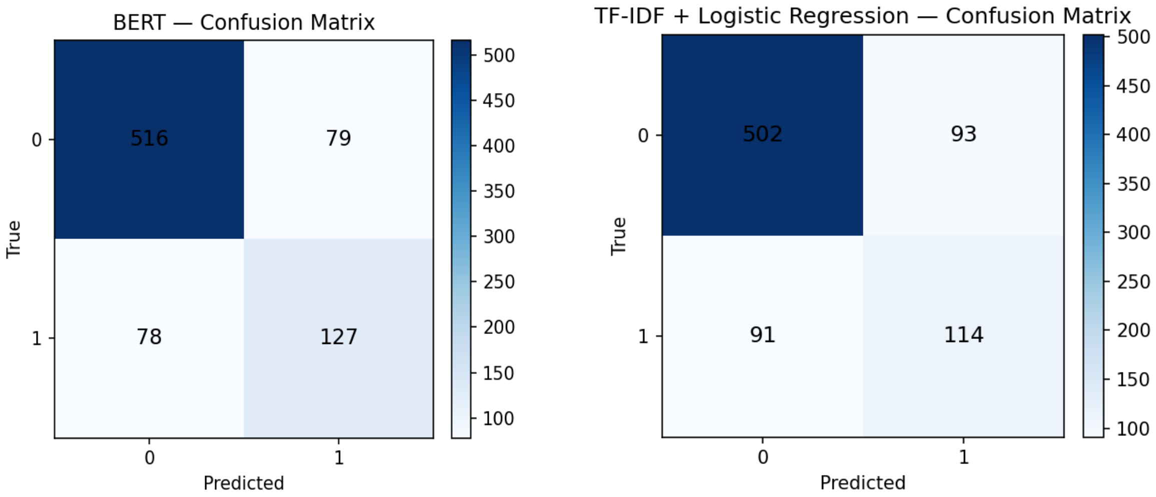

Figure 2. Confusion matrix comparison between Logistic Regression (TF-IDF) and BERT models. The BERT model correctly identifies a larger number of toxic comments (bottom-right cell), indicating improved recall and better separation of toxic versus non-toxic instances.

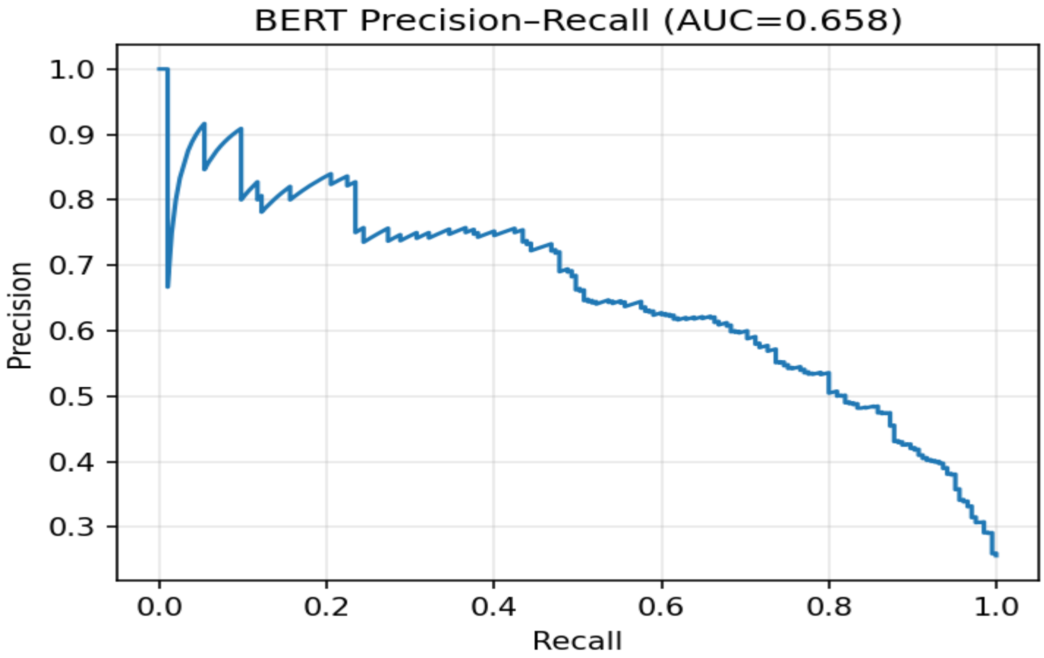

Figure 3. Precision–Recall curve for the fine-tuned BERT model. The area under the curve (AUC = 0.658) demonstrates good balance between precision and recall, confirming the model's effectiveness in detecting toxic comments while limiting false positives.

Key Findings:

1. The fine-tuned BERT model outperforms Logistic Regression across all metrics. It achieves higher accuracy at 80.4 percent compared to 77.0 percent, demonstrates better recall at 62.0 percent versus 55.6 percent, and produces a stronger F1-score of 61.8 percent. This makes BERT the more effective model for toxicity detection in this dataset.

2. Logistic Regression performs respectably as a baseline, with balanced precision and recall around 55 percent. However, its F1-score remains notably lower than BERT's, reflecting its limitations in capturing nuanced toxicity.

3. BERT, while stronger overall, still has room for improvement. Larger variants such as bert-large-uncased or extended training could further boost performance. The model also shows occasional false positives, suggesting that additional fine-tuning or ensemble methods could help balance sensitivity and specificity.

4. Conclusion

This project highlights the superiority of deep contextual models over traditional baselines in toxic comment classification. The Logistic Regression model with TF-IDF features achieved 77.0 percent accuracy with balanced precision and recall around 55 percent, providing a solid benchmark but showing clear limitations in capturing nuanced context. In comparison, the fine-tuned BERT model reached 80.4 percent accuracy, with precision of 61.7 percent, recall of 62.0 percent, and an F1-score of 61.8 percent. These results demonstrate BERT's ability to better balance precision and recall, making it more suitable for real-world moderation tasks.

Future work can extend this project by using larger transformer variants such as BERT-large, applying hyperparameter optimization, exploring ensemble models for robustness, and expanding datasets for greater generalizability.

Overall, this project demonstrates both the promise and the trade-offs of modern NLP: while transformer-based models demand greater computational resources, they offer substantial improvements in recall and balanced detection, making them powerful tools for addressing the complex problem of online toxicity.

For full reproducibility, the complete implementation of this project, including the code, preprocessing pipeline, model training, evaluation, and supplemental materials, is available on GitHub:

Github Repository

References

- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.
- Indiana University, Social Media Mining Course Material, 2025