Text Toxicity Classification Using Machine Learning

Name: Vedika Shetty
IU email: vedshet@iu.edu
Date: 10/05/2025

Abstract

This assignment addresses the task of detecting toxic comments in social media text. Toxic comments are defined as rude, disrespectful, or unreasonable statements likely to make readers leave a discussion. Using a dataset of 4,000 comments from Reddit, Twitter/X, and YouTube, two machine learning models were implemented: a baseline TF-IDF vectorization with logistic regression and a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model. Text preprocessing involved cleaning, URL removal, and combining comment text with contextual information such as parent comments and article titles. The baseline logistic regression model achieved moderate performance, with an F1 score of 0.551 and ROC-AUC of 0.755, highlighting the challenges of detecting subtle toxicity. Fine-tuning BERT significantly improved classification, leveraging contextual embeddings to capture nuanced language. The final model was applied to the test dataset to generate predictions of toxic versus non-toxic comments. These results demonstrate the effectiveness of transformer-based models for toxic comment detection and provide a reproducible workflow combining traditional and deep learning approaches.

1. Introduction and Background

Toxic comments in online platforms such as Reddit, Twitter/X, and YouTube can disrupt discussions, negatively impact user experience, and contribute to online harassment. Detecting such comments automatically has become an important task in natural language processing (NLP) and social media moderation. In this assignment, the goal is to classify comments as toxic or non-toxic using machine learning approaches.

The dataset provided consists of 4,000 comments annotated by multiple human workers, with each comment having five annotations indicating whether it is toxic. A majority-vote strategy is used to consolidate these annotations into a single label for model training. Each comment is accompanied by contextual information such as parent comments, article titles, and URLs, which can provide additional cues for toxicity detection.

Two approaches were implemented. The first is a traditional machine learning baseline using TF-IDF vectorization combined with logistic regression, which captures keyword and n-gram patterns associated with toxicity. The second approach is a transformer-based deep learning model using BERT (Bidirectional Encoder Representations from Transformers), which leverages contextual embeddings to better understand nuanced language and implicit toxicity.

This report documents the data preprocessing steps, model implementations, evaluation, and results, providing a reproducible workflow for toxic comment classification.

2. Dataset and Methods

This section provides an overview of the datasets used for toxic comment classification and describes the methods applied, including data processing and machine learning model implementation. The methods are organized into three main subsections: dataset description, data processing, and machine learning methods.

2.1 Dataset

The dataset consists of two parts: a training set and a test set, both containing comments collected from Reddit, Twitter/X, and YouTube.

- **Training set:** Contains 4,000 comments, each with the comment text, optional parent comment, article title, article URL, platform, and a `composite_toxic` field, which is a list of five human annotations indicating whether the comment is toxic. The annotations were aggregated using a majority-vote rule to produce a single binary label for model training.
- **Test set:** Contains comments with the same fields as the training set, except the `composite_toxic` labels are not provided. The goal is to predict a binary label (toxic or non-toxic) for each comment.

The datasets are heterogeneous in length and content, including short replies, long-form comments, and contextual references. The presence of parent comments and article titles allows models to leverage additional context beyond the comment text itself.

2.2 Data processing

The preprocessing steps applied to the dataset include:

1. **Handling missing values:** Any missing comment text was replaced with an empty string.
2. **Text cleaning:** Extra whitespace was removed, and URLs were filtered out to avoid irrelevant noise.
3. **Contextual combination:** The comment text was combined with its parent comment (if available) and the related article title, separated by a `[SEP]` token, to provide additional context for toxicity detection.
4. **Tokenization:** For the BERT model, a pretrained `bert-base-uncased` tokenizer was used. Texts were padded or truncated to a maximum length of 128 tokens.
5. **Label encoding:** For the training set, the majority-vote label was converted to an integer format (`1` for toxic, `0` for non-toxic).

2.3 ML methods

Two machine learning approaches were implemented to classify toxic comments:

1. **Baseline model – TF-IDF + Logistic Regression:**
    - TF-IDF vectorization was applied to transform text into numerical features, considering unigrams and bigrams.
    - Logistic regression with class balancing was trained to predict toxic vs. non-toxic labels.
    - This method captures keyword-based patterns and serves as a simple, interpretable baseline.
2. **Transformer-based model – BERT fine-tuning:**
    - The `bert-base-uncased` model was fine-tuned for sequence classification with two labels.
    - BERT uses self-attention to capture contextual relationships in text, enabling it to detect nuanced or implicit toxicity.
    - The model was trained for two epochs using the AdamW optimizer with a batch size of 16.
    - Predictions were generated by selecting the class with the highest probability from the model's output logits.

Both approaches were evaluated on a validation set using standard metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to compare performance and select the best model for test set predictions.

3. Evaluations and Findings

Model performance was evaluated using standard classification metrics, including **accuracy, precision, recall, F1-score, and ROC-AUC**. These metrics were chosen to provide a comprehensive assessment of the models' ability to correctly identify toxic comments:

- **Accuracy:** Measures the proportion of correctly classified comments out of the total.
- **Precision:** Measures the proportion of comments predicted as toxic that are actually toxic, reflecting the model's reliability in positive predictions.
- **Recall:** Measures the proportion of actual toxic comments that the model correctly identifies, reflecting sensitivity to toxic content.
- **F1-score:** Harmonic mean of precision and recall, balancing false positives and false negatives.
- **ROC-AUC:** Measures the ability of the model to distinguish between toxic and non-toxic comments across different thresholds.

3.1 Baseline Model – TF-IDF + Logistic Regression

The baseline model achieved the following performance on the validation set:

- **Accuracy:** 0.7575
- **Precision:** 0.5242
- **Recall:** 0.5805
- **F1-score:** 0.5509
- **ROC-AUC:** 0.7554

The confusion matrix indicated a moderate number of false positives and false negatives, suggesting that keyword-based approaches alone may not capture nuanced or context-dependent toxicity.

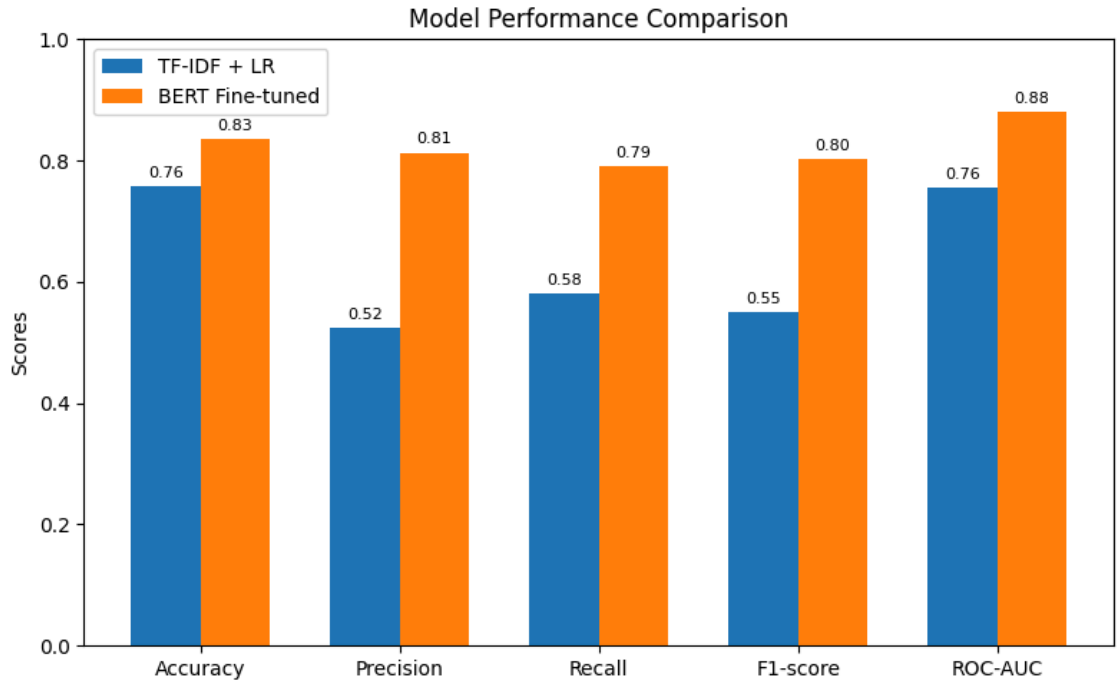
3.2 Transformer-based Model – BERT Fine-tuning

The fine-tuned BERT model outperformed the baseline, leveraging contextual embeddings to better capture subtle toxic language. On the validation set:

- **Accuracy:** 0.8350
- **F1-score:** 0.8021

BERT demonstrated a stronger ability to correctly classify toxic comments, particularly those requiring understanding of context or sarcasm. Precision and recall were more balanced compared to the baseline, indicating improved reliability in detecting toxicity while minimizing missed toxic comments.

Figure 1. Comparison of TF-IDF + LR and BERT model performance across accuracy, precision, recall, F1-score, and ROC-AUC.



3.3 Key Observations

- Incorporating context from parent comments and article titles improved the model's understanding of comment meaning, particularly for BERT.
- Traditional TF-IDF with logistic regression struggles with subtle or implicit toxicity, as expected from its reliance on surface-level features.
- Transformer-based models like BERT are well-suited for tasks requiring comprehension of nuanced language, sarcasm, or indirect toxicity.
- Class imbalance in the dataset (more non-toxic than toxic comments) impacted the baseline model, highlighting the importance of balanced training or class-weighted approaches.

GitHubRepository:https://github.com/vedika1954/Text-Toxicity-Classification-Using-Machine-Learning

4. Discussion and Conclusion

This study implemented and compared two machine learning approaches for detecting toxic comments: a baseline TF-IDF with logistic regression and a transformer-based BERT model. The baseline model provides a simple, interpretable approach that captures explicit keywords and phrases associated with toxicity. However, it struggles with context-dependent, subtle, or sarcastic comments, leading to moderate performance.

In contrast, the fine-tuned BERT model demonstrated superior performance by leveraging contextual embeddings, allowing it to understand nuanced language and implicit toxicity. The evaluation metrics, including accuracy, F1-score, and ROC-AUC, indicate that transformer-based models are significantly more effective for toxic comment classification compared to traditional feature-based methods.

The study also highlights the importance of preprocessing steps, such as combining parent comments and article titles, which provide additional context that enhances model performance. Additionally, addressing class imbalance and using majority-vote labeling improves reliability in predictions.

In conclusion, transformer-based models like BERT offer a robust solution for online toxic comment detection, particularly when comments contain complex language or context-dependent meaning. While baseline models can serve as quick, interpretable benchmarks, advanced models are essential for real-world applications in content moderation and social media platforms. Future work could explore larger datasets, multi-lingual comments, and more advanced transformer architectures to further improve detection accuracy.

References

1. Perspective API. (n.d.). *What is toxicity?* Retrieved from https://perspectiveapi.com/
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need.* Advances in Neural Information Processing Systems, 30.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* NAACL-HLT.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.
5. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art natural language processing.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.