Toxic Comment Classification Using Machine Learning and BERT

Abhir Iyer
abhiyer@iu.edu
02/16/2025

Abstract

Toxic comments on online platforms have become a major concern, affecting user experiences and leading to discussions around content moderation. This project focuses on implementing and evaluating two machine learning models—a Logistic Regression model and a fine-tuned BERT model—to classify whether a given comment is toxic or not. We analyze a dataset consisting of comments from Reddit, Twitter, and YouTube and preprocess them before applying these models. Our findings indicate that BERT significantly outperforms Logistic Regression in terms of both accuracy and recall, making it a better choice for detecting toxic comments.

1. Introduction and Background

Online discussions on social media platforms often contain toxic comments that can be harmful to users. Detecting such content is essential for improving user experiences and enabling better content moderation. While traditional machine learning methods such as Logistic Regression can be used for text classification, they often struggle to capture the nuances of language. Recent advances in Natural Language Processing (NLP) have led to the development of Transformer-based models like BERT, which are more effective at understanding the context of text.

This report describes the implementation of two models:

- Logistic Regression with TF-IDF Features
- Fine-Tuned BERT Model

We compare their effectiveness in detecting toxic comments based on key evaluation metrics, including accuracy, precision, recall, and F1-score.

2. Dataset and Methods

To classify toxic comments, we use a dataset of social media comments and apply machine learning techniques to analyze their toxicity. This section details the dataset characteristics, preprocessing steps, and the machine learning models implemented.

First, we describe the dataset, including its structure and label assignment process. Next, we outline the data preprocessing steps necessary for effective model training. Finally, we explain the machine learning methods used, including Logistic Regression with TF-IDF and a fine-tuned BERT model, highlighting their strengths and limitations.

2.1 Dataset

We use a dataset containing **4,000 comments** collected from Reddit, Twitter, and YouTube. The dataset consists of:

- **Training Set**: Includes comments labeled with **human-annotated toxicity scores**.
- **Test Set**: Contains unlabeled comments for which our models generate predictions.

Each comment is associated with the following attributes:

- `text`: The actual comment content.
- `parent_comment`: The comment being replied to (if applicable).
- `article_title` **&** `article_url`: Context from the linked article (if available).
- `platform`: Source of the comment (Reddit, Twitter, or YouTube).
- `platform_id`: Unique identifier for each comment.
- `composite_toxic`: A list of five human-annotated labels indicating whether the comment is toxic.

To determine the ground truth toxicity label, we applied **majority voting** across the five annotations.

2.2 Data processing

To ensure the dataset is suitable for training, we performed the following data preprocessing steps:

1. Text Cleaning:
    - Converted text to lowercase.
    - Removed special characters, numbers, and extra spaces.
2. Majority Voting for Toxicity Labeling:
    - Assigned a comment as "toxic" if more than 50% of human annotators labeled it as toxic.
3. Train-Test Split:
    - The dataset was split into 80% training and 20% validation to evaluate model performance.
4. TF-IDF Vectorization (For Logistic Regression):
    - Extracted text features using TF-IDF (Term Frequency-Inverse Document Frequency) with bigrams.

2.3 ML methods

Logistic Regression with TF-IDF

- **Model Description**: A traditional text classification model that applies **TF-IDF feature extraction** followed by **Logistic Regression**.

- **Training Setup**:
    - Vectorized text using **TF-IDF with max features = 5,000**.
    - Trained a **Logistic Regression model** with **L2 regularization**.
- **Limitation**:
    - Cannot capture the full **context of words** (e.g., sarcasm, multi-word toxicity).

Fine-Tuned BERT Model

- **Model Description**: A deep learning model based on the **Transformer architecture**, specifically pre-trained **BERT (Bidirectional Encoder Representations from Transformers)**.
- **Training Setup**:
    - Used **"bert-base-uncased"** as the pre-trained model.
    - Fine-tuned on our dataset using **3 epochs** with a **batch size of 8**.
    - **Learning rate**: `2e-5` with **AdamW optimizer**.
    - Used **gradient accumulation** and **mixed-precision training** for efficiency.
- **Strengths**:
    - Understands the **context of words in sentences**.
    - Captures **long-range dependencies** in text.

3. Evaluations and Findings

We evaluated both models using the following metrics:

Accuracy: Measures the overall correctness of predictions.
Precision: Indicates how many predicted toxic comments were actually toxic.
Recall: Measures how many actual toxic comments were correctly detected.
F1-Score: The harmonic mean of precision and recall.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	77.0%	72.3%	16.6%	26.9%
BERT Model	80.6%	63.2%	58.5%	60.8%

Key Findings:

1. BERT significantly outperforms Logistic Regression:

- Higher accuracy (80.6%) compared to Logistic Regression (77.0%).
- Much better recall (58.5%), meaning it detects toxic comments more effectively.
- Balanced F1-score (60.8%), making it a better model overall.

2. Logistic Regression struggles with recall:

- It has high precision (72.3%) but low recall (16.6%).
- This means it misses many toxic comments, leading to an unbalanced detection.

3. Limitations of BERT:

- Training on larger datasets or using bert-large-uncased could further improve results.
- False positives remain an issue, meaning some non-toxic comments are wrongly labeled as toxic.

4. Conclusion

This project demonstrates that **fine-tuning a BERT model significantly improves toxic comment classification** compared to traditional machine learning methods like Logistic Regression. **BERT captures context better, leading to superior recall and F1-scores.**

However, further improvements can be made:

1. **Training on a larger dataset** to enhance generalization.
2. **Using `bert-large-uncased` instead of `bert-base-uncased`**.
3. **Hyperparameter tuning** to fine-tune learning rates and batch sizes.
4. **Applying ensemble models** (combining multiple models for better performance).

Overall, BERT proves to be a powerful model for **real-world toxic comment detection**, helping platforms moderate harmful content effectively.

For full reproducibility, the complete implementation of this project, including code, dataset preprocessing steps, and model training, is available on GitHub: https://github.com/abhiriyer/SMM-Assignment-1

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[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*.