# Assignment 1: Text Analysis

FRENY REJI
fnfren@iu.edu
10/01/2025

**Abstract**

This project focused on building machine learning models to classify whether online comments are toxic or not. We experimented with two approaches: a baseline TF-IDF + Logistic Regression model and a fine-tuned BERT model. The dataset contained 4000 comments from Reddit, Twitter/X, and YouTube with multiple annotators per comment. After preprocessing and majority-vote labeling, we trained and evaluated both models. Results showed that while the TF-IDF baseline achieved 74% accuracy, it struggled on toxic comments (F1 ≈ 0.50). Fine-tuned BERT achieved stronger overall performance, with F1 ≈ 0.61 for toxic comments, showing the value of contextual language models.

## 1. Introduction and Background

Harmful language in digital spaces can be harmful, deter users from participating, and result in increased moderation costs. Detection using an automated approach is not necessarily straightforward, as assessing toxicity can be context-dependent, sarcastic, or slight word choices.

In this assignment, we implemented and compared two models:

- **TF-IDF + Logistic Regression**: a simple linear baseline using bag-of-words style features.

- **Fine-tuned BERT**: a transformer-based language model pre-trained on large text corpora and adapted to this binary classification task.

## 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

The training dataset contained 4000 comments from Reddit, Twitter/X, and YouTube. Each comment had multiple annotations under the `composite_toxic` field, where annotators labeled it as toxic (`true`) or not toxic (`false`). We aggregated these annotations using **majority vote** to assign a single binary label.
The test dataset provided comment text and a `platform_id` for each entry, but no labels. This dataset was used only for generating predictions for submission.

### 2.2 Data processing

- **Cleaning**: removed URLs, extra whitespace, and standardized text formatting.
- **Label aggregation**: converted multiple annotator votes into one binary label (toxic = 1, not toxic = 0).
- **Splitting**: stratified split into 60% training, 20% validation, 20% test to maintain class balance.

### 2.3 ML methods

#### Baseline: TF-IDF + Logistic Regression

- Represented comments as TF-IDF features with unigrams and bigrams (up to 20,000 features).
- Logistic regression classifier with `class_weight='balanced'` to handle class imbalance.

#### Fine-tuned BERT

- Used `bert-base-uncased` tokenizer and model. Input length capped at 128 tokens.
- Fine-tuned for 3 epochs with AdamW optimizer and learning rate 2e-5.
- Evaluation performed at the end of each epoch to monitor overfitting.

## 3. Evaluations and Findings

We evaluated both models using accuracy, precision, recall, and F1-score, focusing especially on the F1-score for the toxic class since accuracy alone can be misleading in imbalanced datasets.

The **TF-IDF + Logistic Regression** baseline reached about 74% accuracy on the test set, with a toxic class F1-score of around 0.50 (precision ≈ 0.49, recall ≈ 0.52). This shows the model

was simple, fast, and interpretable, but it struggled to capture the nuances of context-dependent toxicity.

The **fine-tuned BERT** model achieved higher overall accuracy of about 79% and improved the toxic class F1-score to around 0.61, with precision near 0.60 and recall near 0.62 in its best epoch. These results indicate that BERT could better detect toxic comments while still keeping a balance between precision and recall. However, training beyond the first epoch led to overfitting, where validation performance declined even as training loss decreased.

In summary, ***BERT outperformed the TF-IDF*** baseline by roughly 11 points in toxic class F1-score, demonstrating the advantage of using context-aware transformers for this classification task.

| Model | Accuracy | Precision(Toxic) | Recall(Toxic) | F1(toxic) |
|---|---|---|---|---|
| TF-IDF + LR | 74% | 0.49 | 0.52 | 0.50 |
| BERT | 79% | 0.60 | 0.62 | 0.61 |

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

## 5. References


1. https://www.datacamp.com/blog/what-is-bert-an-intro-to-bert-models

2. https://medium.com/@ryblovartem/text-classification-baseline-with-tf-idf-and-logistic-regression-2591fe162f3b