TOXIC COMMENT CLASSIFICATION

Instructor: Stergios Christodoulidis

Group Members:
Prashant ARYA
Kaushnav ROY

CentraleSupélec

Abstract

The rise of user-generated content across social networks, news portals, and forums has led to an increasing need for automated moderation of toxic or abusive comments. Natural Language Processing (NLP) models have become essential for this task, enabling efficient and scalable detection of harmful content. In this report, we reviewed state-of-the-art approaches in toxic comment classification, focusing on transformer-based models. We evaluate multiple models on a large comment dataset, validate their performance on a separate validation set, and predict results on a test set. Our findings demonstrate promising accuracy, with models achieving over 70% performance across all evaluated algorithms.

Introduction

Given the vast amount of user-generated content, manually moderating toxic comments is impractical. Traditional rule-based filtering systems often fail to capture the complexity of human language, including sarcasm, implicit bias, and evolving slang. To address this challenge, Natural Language Processing (NLP) techniques, particularly deep learning models like transformers, have emerged as powerful tools for automated toxic comment detection. These models can analyze textual content, identify patterns of toxicity, and help platforms enforce community guidelines more effectively. Furthermore, toxic comments often target specific identities, including gender, sexual orientation, religion, and race. In addition to identity-based toxicity, comments can also exhibit varying levels of harm, including severe toxicity, obscene language, threats, insults, identity attacks, and sexually explicit content. Addressing these nuanced aspects is crucial for improving model fairness and mitigating bias in automated moderation systems. Developing robust and accurate NLP models for toxic comment classification is essential for ensuring safer online spaces.

The ultimate goal of this study is to develop NLP models that accurately detect toxic comments while ensuring robustness across diverse demographic subpopulations. A key challenge in automated toxicity detection is that models often exhibit bias, performing well on common toxic expressions but poorly on comments mentioning specific demographic identities. To overcome this, we aim to train models that maximize the worst accuracy metric, ensuring that performance is consistent and unbiased across different demographic groups.

To achieve this, we compare and analyze three transformer-based models, evaluating their ability to detect toxic content across identity groups. We also validate model performance using multiple datasets to ensure generalizability. Selecting the right transformer models is crucial for balancing accuracy, efficiency, and bias mitigation in toxic comment classification. In this study, we experiment and compare the performance of three state-of-the-art transformer models: DistilBERT, ELECTRA, and XLNet.

Dataset

The dataset consists of user-generated text comments along with multiple categorical labels indicating various attributes such as identity groups and toxicity. The dataset classifies comments based on their toxicity and identifies whether they contain identity-based attacks, insults, threats, or obscene content. The dataset (train_x, val_x, test_x) includes features which is a text-based comments and corresponding labels (train_y, val_y) that categorize different aspects of toxicity and identity-based attributes. The dataset contains multiple binary labels such as (a) identity attributes: male, female, black, white, LGBTQ, Christian, Muslim, and other religions, and (b) Toxicity categories: identity_any, severe_toxicity, obscene, threat, insult, identity attack, and sexually explicit.

Methodology

Transformer-based models, including DistilBERT, ELECTRA, and XLNet, leverage deep neural networks with self-attention mechanisms to process and understand text in a contextual manner. Each of these models has distinct architectures and training methodologies that impact their efficiency and performance.

DistilBERT is a lightweight, faster version of BERT trained using masked language modeling (MLM) and knowledge distillation, making it computationally efficient while retaining 97% of BERT's accuracy. ELECTRA, on the other hand, introduces Replaced Token Detection (RTD), where a discriminator learns to distinguish real words from replaced ones, leading to better performance with fewer parameters. XLNet improves upon BERT by using Permutation Language Modeling (PLM), which captures bidirectional context without masking tokens, making it more effective for long-text understanding but computationally expensive. Each model presents trade-offs: DistilBERT is the fastest and most efficient, ELECTRA achieves high accuracy with fewer parameters, and XLNet excels in long-range contextual understanding at a higher computational cost.

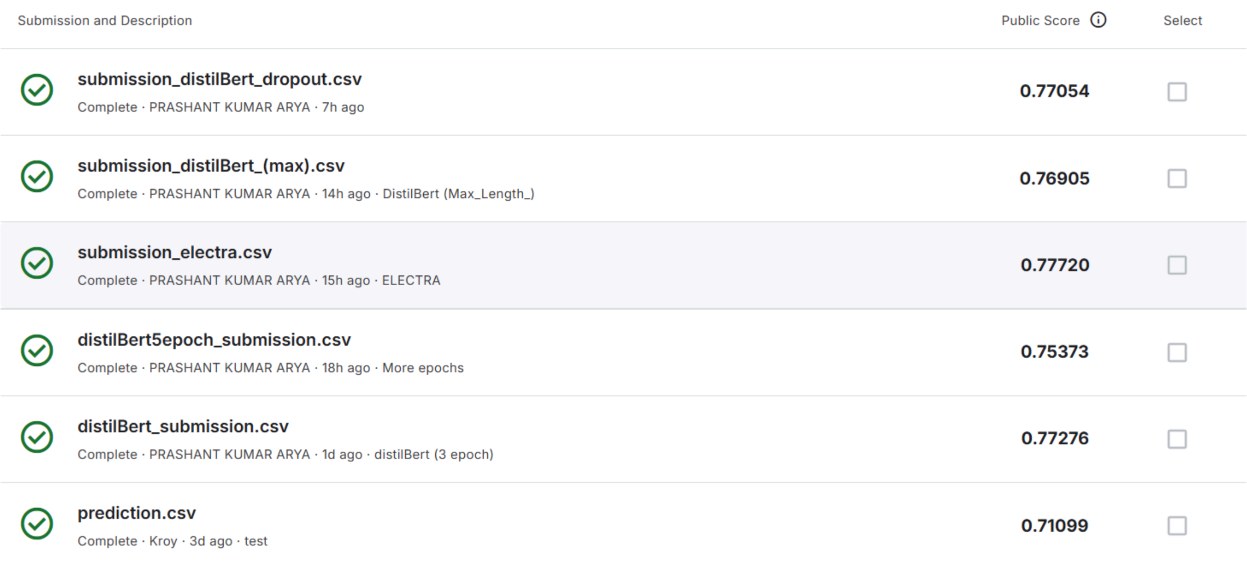
We fine-tuned DistilBERT, ELECTRA, and XLNet using Hugging Face's Transformers library and PyTorch. We performed tokenization using the respective model tokenizers, applying padding, truncation, and a max length of 256 tokens to standardize input sequences. In addition, we utilized the respective model, fine-tuning it for binary classification with a training setup that included 3 epochs, a batch size of 16 per device for both training and evaluation and an AdamW optimizer with a learning rate of 5 x 10^{5}. To stabilize training, we employed 500 warmup steps and applied a weight decay of 0.01 to reduce overfitting. Furthermore, to ensure fair evaluation, we focused on worst-group accuracy, assessing model performance across different identity groups to mitigate bias in classification results.

Results

	Worst Group Accuracy (Validation Set)	F1 Score (Test Set)
DistilBERT	0.7742	0.77276
ELECTRA	0.7706	0.77720
XLNet	Below baseline (<0.7)	Below baseline (<0.7)

The evaluation of DistilBERT, ELECTRA, and XLNet revealed distinct performance characteristics aligned with each model's architecture. ELECTRA emerged as the top performer, achieving the highest F1-score (0.77720) and a robust worst-group accuracy (0.7706), indicating strong generalization across diverse identity groups. Its efficient Replaced Token Detection mechanism contributed to the fastest training time among the models. DistilBERT followed closely, with a competitive F1-score (0.77276) and the highest worst-group accuracy (0.7742), showcasing reliable performance and fairness, while maintaining computational efficiency. In contrast, XLNet struggled with the dataset, performing below baseline (<0.7) in both metrics and proving computationally expensive due to its permutation-based training method. These results suggest that ELECTRA is the most suitable model for tasks requiring balanced performance and efficiency, whereas DistilBERT offers a strong alternative when resource constraints are paramount. XLNet, despite its theoretical advantages in handling complex dependencies, did not align well with the task's requirements, likely due to overfitting or misalignment with the dataset's characteristics.

Submissions



	Submission and Description	Public Score ⓘ	Select
✓	**submission_distilBert_dropout.csv** Complete · PRASHANT KUMAR ARYA · 7h ago	**0.77054**	□
✓	**submission_distilBert_(max).csv** Complete · PRASHANT KUMAR ARYA · 14h ago · DistilBert (Max_Length_)	**0.76905**	□
✓	**submission_electra.csv** Complete · PRASHANT KUMAR ARYA · 15h ago · ELECTRA	**0.77720**	□
✓	**distilBert5epoch_submission.csv** Complete · PRASHANT KUMAR ARYA · 18h ago · More epochs	**0.75373**	□
✓	**distilBert_submission.csv** Complete · PRASHANT KUMAR ARYA · 1d ago · distilBert (3 epoch)	**0.77276**	□
✓	**prediction.csv** Complete · Kroy · 3d ago · test	**0.71099**	□