Toxic Comment Classification with DistilBERT

Byrnes Mulumbeni

Abstract

This project focuses on identifying **toxic and harmful online comments** using a multi-label classification approach. Leveraging **DistilBERT**, a transformer-based LLM, the goal is to fine-tune the model on the **Jigsaw Toxic Comment Classification dataset** to detect categories like “toxic,” “insult,” and “threat.” The project highlights the role of large language models in promoting safer online spaces through automated content moderation.

Introduction

Toxic language on the internet is a growing concern for online communities. Detecting and filtering offensive comments is a challenging NLP problem due to the subtlety and diversity of toxic expressions. This project applies **transfer learning** via DistilBERT to identify multiple types of toxicity in user comments.

Motivation

Why now?

Online platforms increasingly rely on AI to moderate content at scale. Pretrained LLMs offer robust understanding of language, even in challenging and nuanced cases, without needing millions of training examples.

What makes it feasible?

Pretrained transformer models via Hugging Face
Publicly available labeled dataset (Jigsaw)
Colab + GPU make training large models efficient

Problem Formulation

Type: Multi-label text classification (6 classes: toxic, severe toxic, obscene, threat, insult, identity hate)
Dataset: Jigsaw Toxic Comment Classification Challenge (Kaggle)
Libraries/Tools: Python, Hugging Face Transformers, PyTorch, scikit-learn
Model: Fine-tuned DistilBERT
Metrics: Micro & macro F1-score, ROC-AUC, Precision/Recall per label
Validation Strategy: 80/20 train-validation split, stratified by toxicity presence

Workplan

Explore the data: Visualize label distribution, comment lengths
Preprocess:
- Clean and tokenize text
- Encode multi-label targets

Modeling:
- Fine-tune DistilBERT on multi-label objective
- Apply sigmoid activation for multi-label outputs

Evaluate:
- Precision, Recall, F1-scores per class
- ROC curves, confusion matrices

Package & Present:
- Finalized Colab notebook
- Clear documentation and sample predictions