GEN AI PROJECT

Name:- Mithun R
SRN:- PES2UG23CS341

Section:- F

Cyberbullying Detector

- **Goal**: Flag toxic comments on a forum.
- **Tech**: text-classification (toxicity model).

Abstract:-

This project implements a Cyberbullying Detector using Natural Language Processing (NLP) to identify toxic and harmful comments in online forums. The system uses a pre-trained BERT-based toxicity classification model to analyze text input and flag potentially harmful content such as bullying, harassment, or offensive language. The model classifies comments as either "toxic" or "safe" with a confidence score, helping moderators filter out harmful content and create safer online communities.

What I Understood C

yberbullying and online harassment are serious issues in digital communities. Traditional keyword-based filtering systems are inadequate as they can be easily bypassed. Machine learning models, particularly transformer-based models like BERT, can understand context and nuance in language, making them more effective at detecting toxic behavior. The `unitary/toxic-bert` model is specifically trained to identify various forms of toxicity including: - Insults and personal attacks - Threats and harassment - Hate speech - Obscene language

What I Built

I developed a simple yet effective toxicity detection system using the Hugging Face transformers library.

The system:

1. **Loads Pre-trained Model**: Uses the `unitary/toxic-bert` model which is fine-tuned specifically for toxicity detection

2. **Text Classification Pipeline**: Implements a text-classification pipeline that processes input text

3. **Toxicity Analysis**: Analyzes comments and returns: - Classification (toxic/safe) - Confidence score (0-100%)

4. **Interactive Testing**: Provides both batch testing and interactive mode for real-time comment checking

Technical Implementation

- **Library**: Transformers (Hugging Face)

- **Model**: unitary/toxic-bert

- Task: Binary text classification

- **Input**: Text comments/messages

- **Output**: Label (toxic/safe) + Confidence score

Key Features

- Real-time toxicity detection

- Confidence scores for each prediction

- Simple and easy-to-use interface

- Can be integrated into forums, chat applications, or social media platforms

```
Loading model...
Loading weights: 100% ████████████████████████████████ 201/201 [00:00<00:00, 706.05it/s, Materializing param=classifier.weight]
BertForSequenceClassification LOAD REPORT from: unitary/toxic-bert
Key                          | Status     |  |
-----------------------------+------------+--+-
bert.embeddings.position_ids | UNEXPECTED |  |

Notes:
- UNEXPECTED    :can be ignored when loading from different task/architecture; not ok if you expect identical arch.
✅ Ready!

============================================================
🛡️ CYBERBULLYING DETECTOR
============================================================

🚨 TOXIC (99.1%)
Comment: "You're so stupid and ugly!"

🚨 TOXIC (0.8%)
Comment: "Great job on your project!"

🚨 TOXIC (89.1%)
Comment: "Nobody likes you, loser"

🚨 TOXIC (0.1%)
Comment: "I love your work, keep it up!"

🚨 TOXIC (89.4%)
Comment: "You should just quit, you're terrible"

============================================================

Enter comment to check (or 'quit'):  "You are soo dumb"
🚨 TOXIC - Confidence: 93.5%
⚠️ This comment may contain bullying/harassment

Enter comment to check (or 'quit'):  quit
```