PROJECT PROPOSAL

Detecting Toxic Comments Using Deep Learning and Python

Team Members:
KUSHAL REDDY SINGARAM
ZAHID
SAI SARAN RANGISETTI

Problem Statement

The internet can be a harsh place filled with toxic comments, including hate speech, threats, and harassment. Such negativity discourages healthy discussions and affects online communities. Traditional moderation methods are slow and ineffective, requiring a scalable **AI-powered solution** to detect and filter toxic content automatically.

Business Need

An **accurate and real-time toxicity detection system** is essential for social media platforms, forums, and news websites to:

- **Protect users** from online harassment.
- **Reduce manual moderation efforts** and enhance efficiency.
- **Improve engagement** by fostering positive interactions.
- **Ensure compliance** with content guidelines and regulations.

Project Objective

- Develop a **Deep Neural Network (DNN)-based model** to classify comments as toxic or non-toxic.
- Utilize **cables and Radio apps** for **seamless integration** and real-time predictions.
- Allow **custom dataset substitution** for platform-specific needs.
- Provide an interactive system to analyze and predict toxicity from raw text.

Techniques & Methodology

Component	Details
Model Development	• Deep Neural Network (DNN): Multi-layer architecture to capture complex language patterns. • Cables Integration: Connect deep learning components for real-time predictions. • Radio Apps: Enable model deployment and user-friendly interaction.
Frameworks & Tools	• Deep Learning: TensorFlow/Keras for model training. • NLP Libraries: NLTK, spaCy for preprocessing. • Deployment Tools: FastAPI/Flask for real-time API integration.
Training & Hyperparameter Tuning	• Optimize learning rate, batch size, and activation functions. • Use dropout and batch normalization to prevent overfitting.
Evaluation Metrics	• Classification Metrics: Accuracy, F1-score, Precision, Recall, ROC-AUC. • Bias Detection: Ensure fair predictions across diverse user comments.
Interpretation & Insights	• Visualize prediction confidence with SHAP/LIME. • Generate real-time toxicity scores from raw text inputs.

Dataset Name & Characteristics

- **Dataset:** Jigsaw Toxic Comment Classification
- **Characteristics:**
  - **Contains real user comments** from Wikipedia discussions.
  - **Labeled into multiple toxicity categories:** toxic, severe toxic, obscene, threat, insult, identity hate.
  - **Multilabel Classification Problem:** A comment can belong to multiple categories.
  - **Large-scale dataset:** Over 150,000 labeled comments.