Project Synopsis

Title:

Multi-Label Text Classification for Toxic Comment Detection using Machine Learning.

Objective:

The primary objective of this project is to develop a robust and efficient machine learning-based system that can classify online comments into multiple predefined toxicity categories. Unlike traditional single-label classification tasks, this project tackles a multi-label classification problem where each comment may belong to more than one toxicity class. The system aims to automate the detection of offensive and harmful online content, thus contributing to safer digital communication environments.

Problem Statement:

In the age of digital communication and social media, user-generated content is often susceptible to offensive, harmful, and toxic language. Managing such content manually is impractical due to the volume of data generated every second. Furthermore, toxic comments can possess multiple harmful attributes simultaneously, making detection even more challenging.

The task involves:

- - Detecting various types of toxic behavior: toxic, severe toxic, obscene, threat, insult, and identity hate.
- - Classifying comments where multiple labels may be applicable.
- - Delivering accurate and explainable results using interpretable machine learning models.

Methodology:

1. 1. Data Acquisition:

- The dataset used is the publicly available Jigsaw Toxic Comment Classification dataset from Kaggle.

2. 2. Data Preprocessing:

- Converting text to lowercase

- Removing special characters, digits, unnecessary white spaces

- Removing stopwords using NLTK

- Stemming using Snowball Stemmer

3. 3. Feature Engineering:

- TF-IDF vectorization with unigrams and bigrams

- Max features set to 10,000

4. 4. Model Implementation:

- Logistic Regression using One-vs-Rest strategy

- Multinomial Naive Bayes using One-vs-Rest strategy

5. 5. Model Evaluation:

- Metrics include classification reports, ROC curves, and label distribution visualizations

Tools & Technologies:

Python 3, Pandas, NumPy, NLTK, Scikit-learn, Matplotlib, Seaborn

Key Results:

- - Multi-label classification pipeline for detecting various types of toxicity
- - Logistic Regression and Naive Bayes demonstrated effective baseline performance
- - ROC Curves for interpretability

Expected Outcomes:

- - A functional, reproducible machine learning pipeline for toxic comment detection
- - Model performance metrics for each category
- - Visualizations for performance reporting

Future Scope:

- - Integrating deep learning models (LSTM, GRU, BERT)
- - Deploying trained models via REST API or web application
- - Multilingual support and explainability modules

Conclusion:

This project provides a scalable machine learning solution for moderating toxic online content, offering a foundation for future advancements in deep learning-based text classification.