Hate Speech and Offensive Language Detection Using NLP

Abstract:
Hate speech and offensive language in online platforms are major issues that lead to harmful social interactions and platform abuse. This project aims to build an **NLP-based system for detecting hate speech and offensive language** in user-generated content, such as social media comments, forum posts, and reviews.

Methodology:
The system will preprocess text using **tokenization**, **lemmatization**, and **stopword removal**. The core detection will be based on **sentiment analysis**, **toxic language identification**, and **contextual analysis** using models like **BERT** or **RoBERTa**. The system will be trained on labeled datasets containing various categories of offensive language, including racial slurs, gender-based insults, and bullying. **Supervised learning algorithms** like **SVM**, **Random Forest**, and **neural networks** will be used to classify content into categories such as hate speech, offensive, or neutral.

Outcome:
The expected outcome is a real-time hate speech detection system that can automatically flag or filter harmful content on social media, online communities, and review platforms. By preventing the spread of offensive language, the system will contribute to healthier online interactions and compliance with platform moderation policies.