NLP Project

Hate Speech Detection Project Submission

Individual Project

- Name: Riya Gaur
- **Email:** riyagaur1299@gmail.com
- **Country:** USA
- **Specialization:** NLP (Natural Language Processing)

Problem Description

Hate speech is any form of communication, whether verbal, written, or behavioral, that attacks or discriminates against individuals or groups based on attributes like religion, ethnicity, nationality, race, color, sex, or other identity factors. It has significant social and ethical implications, particularly in the digital space. This project focuses on developing an NLP-based system to detect hate speech in textual data, specifically Twitter tweets, to promote safer and more inclusive online environments.

Business Understanding

Detecting hate speech is a critical task for organizations like social media platforms, content moderators, and regulatory bodies. By automating this task, businesses can:

- Enhance user safety and experience.
- Identify and remove harmful content efficiently.
- Uphold compliance with legal and ethical standards.
- Foster inclusivity and discourage toxic online behavior.

This project aims to deliver a robust hate speech detection model that can classify tweets as containing hate speech or not, enabling scalable and reliable moderation efforts.

Project Lifecycle and Deadline

1. **Problem Understanding**
2. **Data Collection and Cleaning**
3. **Data Representation and Exploration**
4. **Model Building and Training**
5. **Performance Evaluation**

6. **Deployment and Inference**
   Deadline: Feb 28

Data Intake Report

Dataset:

- **Source:** Twitter dataset containing tweets labeled for hate speech detection.
- **Attributes:**
  - `label`: Binary classification (0 for non-hate speech, 1 for hate speech).
  - `text_format`: Original tweets with potential noise.

Data Overview:

- Number of records:
- Distribution of labels:
- Data characteristics: High noise, presence of emojis, slang, and informal language.

Preprocessing Steps:

1. Remove special characters, hashtags, and URLs.
2. Convert text to lowercase.
3. Tokenize and remove stopwords.
4. Perform stemming and lemmatization.