Hate Speech Detection - Assignment Report

Individual Project
Name: Riya Gaur
Email: riyagaur1299@gmail.com
Specialization: NLP (Natural Language Processing)

Introduction

Hate speech detection is a crucial Natural Language Processing (NLP) task that helps identify and filter out harmful content in online platforms. This project focuses on cleaning and preprocessing a dataset of tweets to improve the accuracy of a hate speech classification model.

2. Data Description

- **Dataset Name:** `train_E6oV3lV.csv`
- **Number of Records:** 31,962 tweets
- **Columns:**
  - `id` (Unique identifier for each tweet)
  - `label` (0 = Non-hate speech, 1 = Hate speech)
  - `tweet` (Original text content)

3. Data Preprocessing & Cleansing

3.1 Handling Missing Values

- Checked for missing values in the dataset.
- No missing values were found in the dataset.
- If missing values existed, they would have been replaced using techniques like:
  - Filling missing `tweet` text with "No text available."
  - Dropping rows with missing labels.

3.2 Removing Outliers

- Applied **Interquartile Range (IQR) Method** to detect and remove extreme text lengths.
- Applied **Z-score filtering** to remove tweets with unusually high word counts.
- This ensures that the dataset is free from extreme variations in text length.

3.3 Text Cleaning

- Applied regular expressions to:
  - Remove **@mentions**.
  - Remove **URLs**.
  - Remove **special characters and numbers**.
  - Convert text to **lowercase**.

3.4 Tokenization, Lemmatization & Stopword Removal

- Used **spaCy NLP model to:
  - Tokenize words into meaningful units.
  - Remove **stopwords** (common words that don't add value, e.g., "the", "is").
  - Convert words to their **lemmatized (root) form**.
- Saved the cleaned dataset as `processed_train.csv` for further analysis.

4. Exploratory Data Analysis (EDA)

4.1 Class Distribution

- **92.98%** of the tweets were **non-hate speech** (`label = 0`).
- **7.02%** of the tweets were **hate speech** (`label = 1`).
- The dataset is **highly imbalanced**, which may require oversampling (SMOTE) or weighting strategies during model training.

4.2 Word Frequency Analysis

- Most frequent words were extracted from hate speech and non-hate speech categories separately.
- Visualization techniques like **word clouds** and **bar plots** were used.

5. Feature Engineering - NLP Featurization

- Applied **TF-IDF Vectorization**:
  - Converted textual data into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF).
  - Limited vocabulary to **5000 most important words**.
- Applied **Count Vectorization (Bag of Words)**:
  - Used CountVectorizer to transform text into word frequency features.
- These transformations prepare the dataset for machine learning models.

6. Results & Conclusion

- The dataset was successfully cleaned, preprocessed, and transformed using NLP techniques.
- Outliers were removed using **IQR and Z-score methods**.
- Featurization using **TF-IDF and CountVectorizer** was implemented to convert text data into numerical form.
- The dataset is now well-prepared for further modeling and classification.