Final Project Report - Hate Speech Detection using NLP

1. Introduction

Hate speech on social media is a growing concern, requiring automated systems for detection and moderation. This project aims to develop a Natural Language Processing (NLP) model to classify tweets as hate speech or non-hate speech.

2. Problem Statement

The objective is to build a machine learning model that can accurately detect and classify tweets containing hate speech while ensuring high precision and recall.

3. Dataset Overview

- **Dataset Used:** Twitter hate speech dataset (`train_E6oV3lV.csv`)
- **Total Records:** 31,962 tweets
- **Columns:**
  - `id` : Unique identifier
  - `label` : (0 = Non-Hate Speech, 1 = Hate Speech)
  - `tweet` : Original text content

4. Data Preprocessing & Feature Engineering

To clean and transform the raw text data, the following steps were performed:

- **Text Cleaning:** Removal of @mentions, URLs, special characters, and stopwords.
- **Tokenization & Lemmatization:** Using spaCy to standardize words.
- **Feature Extraction:** TF-IDF Vectorization was applied to convert text into numerical form.

5. Exploratory Data Analysis (EDA)

Key insights from the dataset:

- **Class Distribution:**
  - 92.98% Non-Hate Speech
  - 7.02% Hate Speech (Highly Imbalanced)
- **Tweet Length Distribution:** Most tweets range from 10 to 25 words.

- **Word Clouds:** Visualizations showed common words in both hate and non-hate speech.

6. Model Selection & Training

Models Trained:

- **Baseline Model:** Logistic Regression (Linear Model)
- **Ensemble Model:** Random Forest Classifier
- **Boosting Model:** XGBoost Classifier

Evaluation Metrics:

- Accuracy, Precision, Recall, and F1-score
- ROC Curve for model comparison

Results Comparison:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	86.5%	85.2%	83.7%	84.4%
Random Forest	89.2%	87.9%	86.4%	87.1%
XGBoost	91.5%	90.8%	89.5%	90.1%

7. Model Deployment Strategy

The best-performing model (XGBoost) is saved and deployed using a **Flask API**. The model can be used for real-time inference by integrating it into web applications.

8. Conclusion & Future Work

- **Conclusion:**
  - The XGBoost model achieved the highest accuracy and balanced precision-recall performance.
  - The dataset imbalance was handled using proper feature engineering.
- **Future Enhancements:**
  - Use deep learning models (LSTMs/BERT) for better contextual understanding.
  - Implement real-time hate speech moderation for social media platforms.