Hate Speech Detection using Transformers (Deep Learning)

Team Member Details

Group Name: NLP Pioneers

Name: Duc Duong

Email: duongduc@grinnell.edu

Country: United States

College: Grinnell College

Specialization: Data Science & NLP

Problem Description

Hate speech refers to written, verbal, or behavioral communication that attacks or discriminates against a person or group based on religion, ethnicity, nationality, race, gender, or identity factors. The goal of this project is to detect hate speech in text data (Twitter tweets) using deep learning transformer models.

Business Understanding

Social media platforms face growing challenges in moderating harmful content. Automated detection of hate speech reduces reliance on manual moderation, ensuring safer online communities and regulatory compliance. A transformer-based detection system can provide scalable, accurate, and real-time solutions to support business needs in content moderation, brand safety, and policy compliance.

Project Lifecycle & Deadline

1. **Business Understanding & Data Collection** (Week 1–2)

2. **Exploratory Data Analysis & Preprocessing** (Week 3–4)
3. **Model Development (Transformers)** (Week 5–6)
4. **Model Evaluation & Optimization** (Week 7–8)
5. **Deployment & Documentation** (Week 9–10)

Deadline: 10 weeks from project start

Data Intake Report

Project: Hate Speech Detection using Transformers (Deep Learning)

Dataset: Twitter Hate Speech Dataset (Kaggle)

Report Date: September 21, 2025

Version: 1.0

Data Intake by: Duc Duong

Reviewer: <To be assigned>

Data Storage Location: https://www.kaggle.com/code/dnghngc/twitter-hate-speech-detection-different-model

Tabular Data Details:

Total number of observations	**49,200 (16,130 unique tweets)**
Total number of files	2 (train + test CSV)
Total number of features	5 (id, tweet, label, etc.)
Base format of the file	CSV
Size of the data	~30 MB
Dataset Source	https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech/data

Proposed Approach:

- Deduplicate based on tweet text and id.
- Normalize and clean tweets (remove special characters, hashtags, URLs, mentions).
- Address class imbalance using SMOTE, oversampling, or weighted loss functions.
- Tokenize with Hugging Face Transformers (BERT, RoBERTa, DistilBERT).
- Train transformer models for classification and evaluate using F1-score, precision, recall.