Data Glacier Internship Project
Batch LISUM36: 30 July – 30 Oct 24
Project: Advance NLP: Hate Speech detection using Transformers (Deep Learning) - Group Project

Team:
Team Name: Team Trailblazers

Members:

Team member one:	Team member two:
Michael Udonna Egbuzobi egbuzobi.michael@gmail.com United Kingdom University of Wolverhampton Data Science	**Nweke Nonye** nonyenweke22@gmail.com United Kingdom University of Wolverhampton Data Science

Problem Description:

Hate speech is a form of communication that uses derogatory language to attack or discriminate against individuals based on aspects like religion, ethnicity, nationality, race, colour, ancestry, or other identity factors. Detecting hate speech online is crucial for maintaining healthy social interactions, particularly on platforms like Twitter, where information spreads quickly. The aim of this project is to develop an advanced hate speech detection model using transformer-based deep learning architectures. The model will classify text (tweets) into hate speech or non-hate speech (binary classification).

Github links

Link to repository: https://github.com/UdonnaM/Advance-NLP-Hate-Speech-detection-using-Transformers-Deep-Learning-

Link to EDA ipynb: https://github.com/UdonnaM/Advance-NLP-Hate-Speech-detection-using-Transformers-Deep-Learning-/blob/main/TWEET%20NLP.ipynb

Exploratory Data Analysis and Recommendations

Exploratory Data Analysis (EDA)

- The complete visualizations and analysis can be found in the accompanying Jupyter Notebook file, **TWEET NLP.ipynb**.

- Two datasets were utilized for this project: **train** and **test** datasets.
    - ✓ The **train** dataset comprises 31,962 rows and 3 columns.
    - ✓ The **test** dataset consists of 17,197 rows and 2 columns.

- **Imbalanced Data**: The distribution of the training data is imbalanced, with significantly more non-hate speech tweets than hate speech tweets.

- **Word Cloud Analysis**:
    - ✓ **Hate Speech Keywords**: Common terms include "trump", "black", "racist", "liptard", and "amp".
    - ✓ **Non-Hate Speech Keywords**: Terms such as "love", "happy", "smile", and "day" were frequently found in non-hate speech tweets.

- **Correlation Analysis**:
    - ✓ A low correlation (0.051) was found between tweet length and hate speech, indicating that tweet length does not significantly predict the presence of hate speech.

Model Training Progress

Over the course of three epochs, the model demonstrated steady improvements, though some concerns about overfitting arose:

- **Epoch 1**:
    - ✓ Training Loss: 0.7092
    - ✓ Validation Loss: 0.5455
    - ✓ Observation: The model started learning patterns, with the validation loss being lower than the training loss.
- **Epoch 2**:
    - ✓ Training Loss: 0.4515
    - ✓ Validation Loss: 0.3762
    - ✓ Observation: Both training and validation losses dropped significantly, indicating improved performance.
- **Epoch 3**:
    - ✓ Training Loss: 0.4281
    - ✓ Validation Loss: 0.6099
    - ✓ Observation: The increase in validation loss suggests overfitting, where the model started memorizing the training data rather than generalizing well to unseen data.

Final Recommendations

- **Address Class Imbalance**: The dataset is heavily skewed toward non-hate speech tweets. While class weights were applied to mitigate this, further actions are recommended:
    - ✓ **Data Augmentation**: Synthetic data generation or collecting additional hate speech examples could further improve model performance.

- **Leverage Advanced NLP Techniques**: Hate speech tweets tend to contain aggressive or inflammatory language. Implementing advanced methods like word embeddings or attention mechanisms may enhance model accuracy.

- **Overfitting Mitigation**: The rise in validation loss during Epoch 3 suggests overfitting. Techniques such as **early stopping**, **dropout**, and **regularization** should be considered to avoid overfitting in future training iterations.