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).

Week 12: Model Selection and Model Building/Dashboard

Transformers model is my base model based on business requirement. We decided to select other model families (Logistic Regression and XGBoost) for comparison and which is more effective for this business requirement.

Model Comparison for Hate Speech Detection

i. Transformers Model

- **Non-hate Class (0)**: The model achieved **precision** of **0.98**, **recall** of **0.99**, and an **F1-score** of **0.98** for the non-hate class, indicating that it correctly identifies non-hate tweets with very high accuracy.
- **Hate Class (1)**: Precision for hate speech is **0.79** and recall is **0.71**, resulting in an F1-score of **0.75**. This indicates that while the model is relatively strong at

identifying hate speech, it misses about **29%** of true hate tweets (false negatives).

- **Overall Performance**:
  - **Accuracy**: **0.97**
  - **Macro Avg F1**: **0.87** (balanced between classes)
  - **Weighted Avg F1**: **0.97** (dominated by the majority class)
- **Interpretation**:
  - The **Transformers model** shows the best overall performance, especially for the **non-hate class**, achieving a high F1-score and accuracy.
  - However, it still struggles with the hate speech class, indicating that further balancing or additional fine-tuning could help improve hate speech detection.

ii. Logistic Regression

- **Non-hate Class (0)**: Logistic Regression achieved **precision** of **0.99**, **recall** of **0.93**, and an **F1-score** of **0.96**. This suggests it's highly accurate at identifying non-hate tweets but slightly less so than the Transformers model.
- **Hate Class (1)**: For hate speech, **precision** is **0.48** and **recall** is **0.84**, resulting in an F1-score of **0.61**. Although it identifies a good portion of hate tweets (high recall), it also mislabels a significant number of non-hate tweets as hate (leading to a lower precision).
- **Overall Performance**:
  - **Accuracy**: **0.92**
  - **Macro Avg F1**: **0.78**
  - **Weighted Avg F1**: **0.93**
- **Interpretation**:
  - **Logistic Regression** performs well on the non-hate class, but its lower precision on the hate class suggests it has a higher rate of **false positives** (mislabelling non-hate as hate).
  - The high recall in the hate class (0.84) indicates that this model is more sensitive to detecting hate speech, but it sacrifices precision.

iii. XGBoost Model

- **Non-hate Class (0)**: XGBoost achieved **precision** of **0.98**, **recall** of **0.93**, and an **F1-score** of **0.95**, showing strong performance but slightly lower than Logistic Regression in non-hate detection.
- **Hate Class (1)**: For hate speech, XGBoost shows **precision** of **0.43** and **recall** of **0.72**, resulting in an F1-score of **0.54**. This performance is lower than both the Transformers and Logistic Regression models in terms of correctly identifying hate tweets.
- **Overall Performance**:

    - **Accuracy**: **0.91**
    - **Macro Avg F1**: **0.74**
    - **Weighted Avg F1**: **0.92**
- **Interpretation**:
    - **XGBoost** performs reasonably well overall but struggles most with the hate class, where it has both lower precision and recall. This suggests that XGBoost has more difficulty distinguishing hate speech compared to the other models.
    - While XGBoost maintains accuracy in detecting non-hate tweets, it has a higher error rate when it comes to the minority class (hate speech), which is reflected in its lower F1-score for hate tweets.

Summary and Comparison

1. **Transformers Model**:
    - **Best overall performance** with high accuracy (0.97) and strong F1-scores for both classes.
    - The **most effective model for both non-hate and hate classes**, but still with room for improvement in hate speech detection.
2. **Logistic Regression**:
    - Performs **well on non-hate tweets** and has a high **recall for hate speech**, meaning it catches more hate speech examples than XGBoost.
    - However, **low precision in the hate class** suggests more false positives, where non-hate tweets are misclassified as hate.
3. **XGBoost**:
    - **Strong performance in non-hate detection** but relatively weak for hate speech.
    - **Lowest F1-score for hate class**, indicating it has more difficulty handling the class imbalance and distinguishing hate speech effectively.

Recommendation

- **Transformers** would be the recommended model due to its **highest overall accuracy and F1-scores**, indicating it generalizes best across both classes.
- **Further fine-tuning** or **balancing techniques** may help improve the hate speech detection further.
- **Logistic Regression** is an alternative if computational resources are limited, as it captures more hate speech (higher recall) but at the cost of precision.