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).

Data Understanding

The dataset provided for this project consists of two files: a training dataset and a test dataset. The training dataset contains three columns: id, which uniquely identifies each record; tweet, representing the textual data of the tweets; and label, indicating whether the tweet contains hate speech (1) or not (0). The test dataset consists of two columns: id and tweet, where the labels are absent, and predictions are to be made based on the trained model. This structure allows for supervised learning, with the training data used to build and tune the model, while the test data serves to evaluate its performance on unseen examples.

- **Missing Values**: The datasets do not contain any missing values, ensuring complete data for analysis and model evaluation.

- **Outliers:** No significant outliers are present in the data, allowing for a smooth analysis without the need for outlier treatment.

- **Skewness:** The data exhibits left skewness, with the majority of tweets classified as "not hate speech." This imbalance in distribution may affect model performance and requires normalization.

- **Class Imbalance:** There is a notable imbalance in the datasets, with significantly fewer instances of hate speech compared to non-hate speech. This class imbalance could lead to biased model outcomes and necessitates the use of techniques like resampling or adjusting class weights.

- **Noise in Data:** The datasets consist of short textual content (tweets), which inherently contain various forms of noise. This noise includes informal language, abbreviations, hashtags, mentions, URLs, emojis, and special characters that do not contribute to the analysis. Such elements can obscure the meaningful patterns required for accurate hate speech detection, making it necessary to address these issues before building a reliable model.

Proposed Solutions

i. Addressing Skewness

- **Approach:** We will apply transformations, such as log transformation, to normalize the data distribution if needed.
- **Rationale:** Normalizing the distribution can improve the performance of many machine learning algorithms that assume normally distributed data.

ii. Managing Class Imbalance

- **Approach:** We will utilize techniques such as:
  - **Resampling:** Applying oversampling (e.g., SMOTE) on the minority class (hate speech) to generate synthetic samples, or under sampling the majority class.
  - **Class Weights:** Adjusting the class weights in our model to give more importance to the minority class.
- **Rationale:** These approaches will help the model learn to identify hate speech more effectively, reducing the risk of bias towards the majority class.

iii. Solutions to Data Noise: To mitigate these challenges, several pre-processing techniques will be employed. Tokenization will break the tweets into individual words, while irrelevant components such as stop words, hashtags, mentions, and URLs will be removed. Additionally, emojis and special characters will either be normalized or excluded from the text. By implementing these steps, we aim to clean and standardize the data, ensuring the model can focus on the linguistic patterns essential for detecting hate speech.

Conclusion

This report outlines our understanding of the dataset and the challenges we face in detecting hate speech. By addressing skewness and class imbalance, we aim to build a robust model capable of accurately classifying tweets. Continuous evaluation and refinement of our approaches will ensure we achieve the best possible outcomes for our hate speech detection project.