Data Glacier

Your Deep Learning Partner

# Advance NLP : Hate Speech detection using Transformers

Michael Udonna Egbuzobi and Nweke Nonye

**16-October-2024**

# Agenda

Problem Statement

EDA

Recommended Models for Hate speech detection

Team Members

Link to repository

# Problem Statement

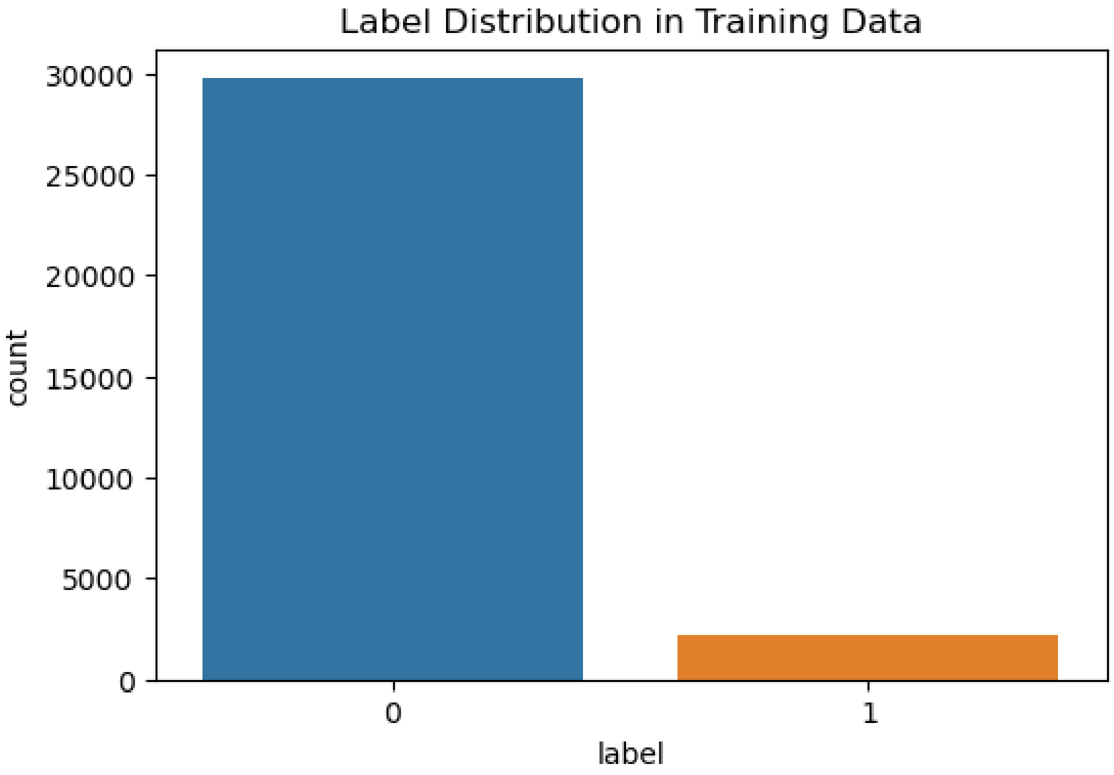
Hate speech is a form of communication that uses derogatory language to attack or discriminate against individuals.

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).

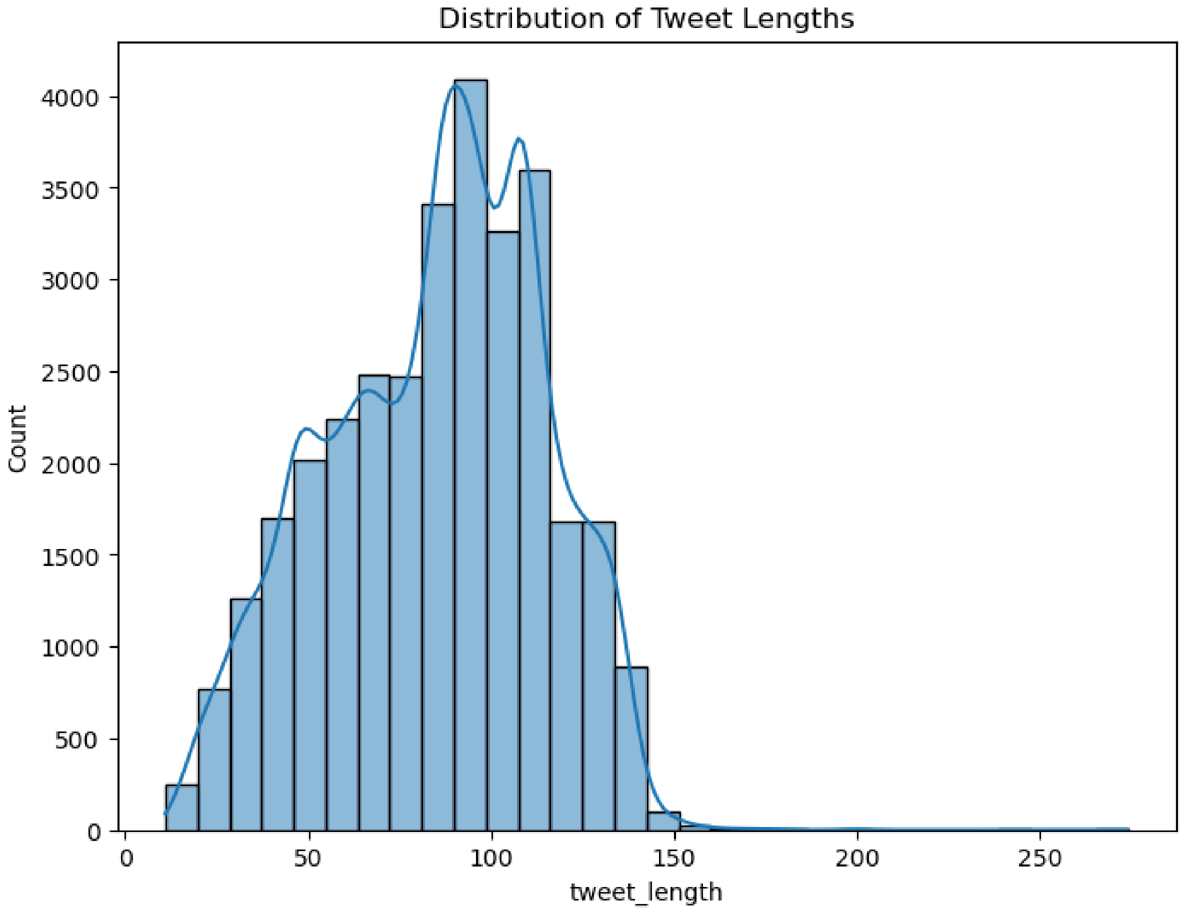
# Exploratory Data Analysis

We initiated our analysis by assessing the distribution of labels and identified an imbalance in the dataset. To address this, we applied class weighting to ensure more equitable model training.



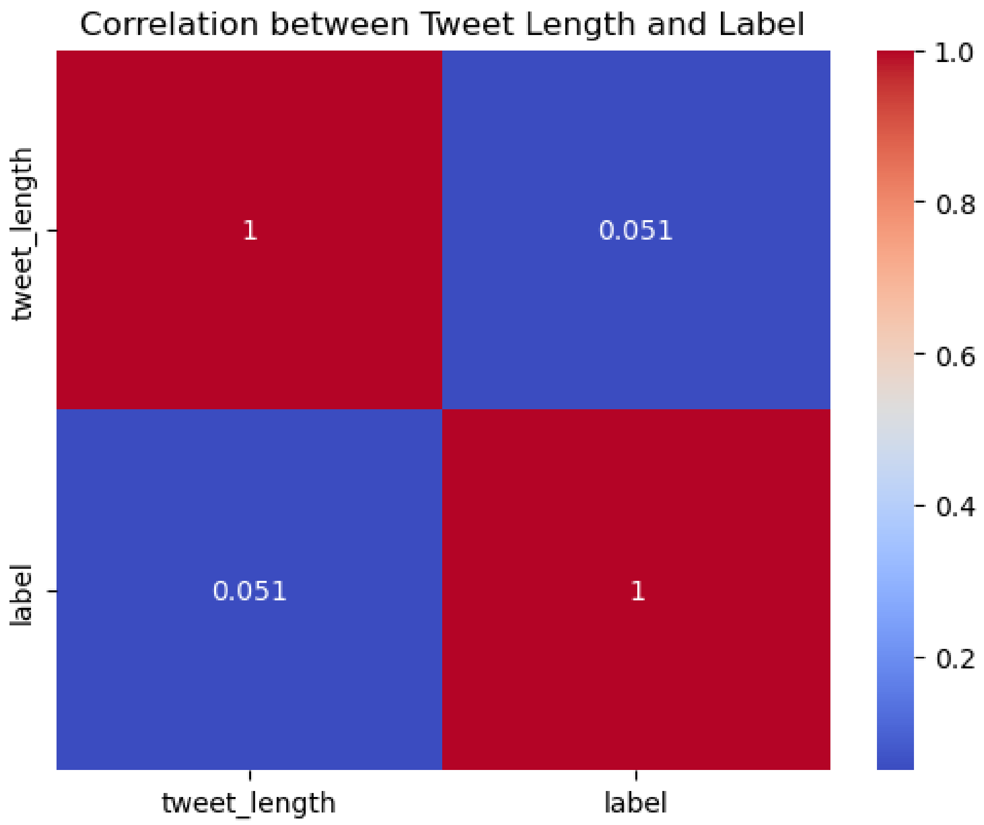
# Exploratory Data Analysis

Next, we analyzed the distribution of tweet lengths and observed a left-skewed pattern, further indicating the presence of data imbalance.



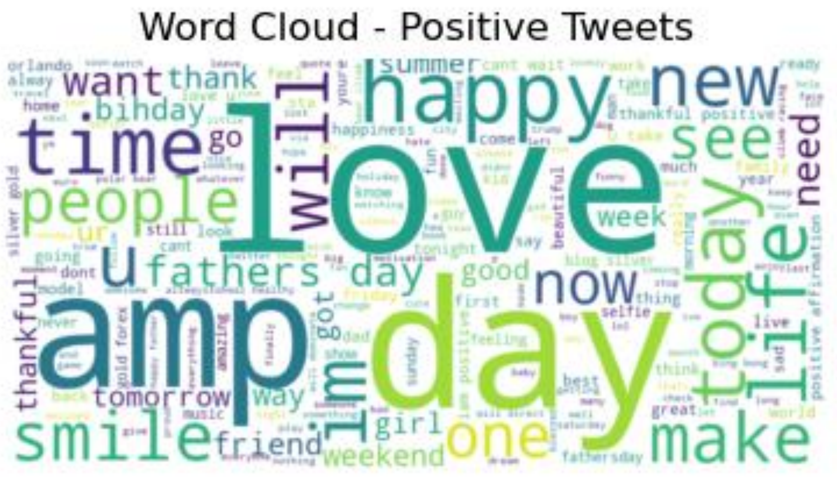
# Exploratory Data Analysis

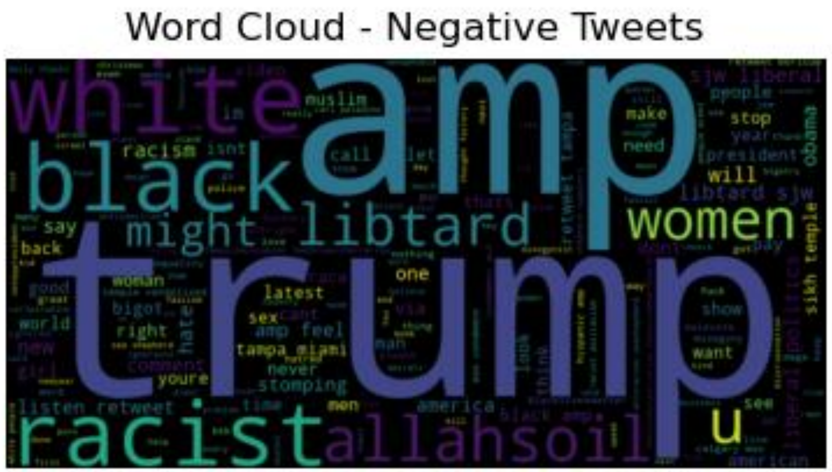
Correlation Analysis: A low correlation (0.051) was found between tweet length and hate speech, indicating that tweet length does not significantly predict the presence of hate speech.



# Exploratory Data Analysis

- **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.


Word Cloud - Positive Tweets


Word Cloud - Negative Tweets

# Recommended Models for Hate Speech Detection

**Model:** Transformer-based models (e.g., BERT)

**Rationale:**

- **Contextual Understanding:** These models excel at capturing the nuances and subtleties of language, which is crucial for accurately identifying hate speech.
- **State-of-the-Art Performance:** Extensive research shows that transformer models consistently outperform traditional approaches in various NLP tasks, making them a powerful choice for our project.
- **Fine-Tuning Flexibility**: By leveraging pre-trained models, we can effectively adapt them to our specific dataset, minimizing the need for large labelled samples while maximizing performance.
- **Addressing Imbalance:** We employed class weighting during training to ensure that our model learns to recognize minority classes effectively, which is essential given the imbalanced nature of our dataset.

**Conclusion:** In summary, transformer-based models offer the robustness and precision needed for reliable hate speech detection in tweets, aligning perfectly with our project goals.

# Team Members: Team Trailblazers

**Team member one:**

Michael Udonna Egbuzobi
egbuzobi.michael@gmail.com
United Kingdom
University of Wolverhampton
Data Science

**Team member two:**

Nweke Nonye
nonyenweke22@gmail.com
United Kingdom
University of Wolverhampton
Data Science

# Link to Repository

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

# Advance NLP : Hate Speech detection using Transformers

Data Glacier

Your Deep Learning Partner

Thank You