Advance NLP : Hate Speech detection using Transformers

Michael Udonna Egbuzobi and Nweke Nonye

31-October-2024

Agenda

- Project Overview
- Problem Description
- Data Source
- Data Preprocessing
- Exploratory Data Analysis
- Model Selection
- Model Training & Evaluation Metrics
- Comparison of Model Performance
- Recommendation
- Conclusion
- Team Members
- Link to repository

Project Overview

Developed an advanced hate speech detection model using transformer-based deep learning, leveraging state-of-the-art NLP techniques.

Addressed the growing prevalence of hate speech on social media, focusing on platforms like Twitter where harmful content impacts social harmony.

Aimed to classify tweets as hate speech or non-hate speech, supporting safer online interactions and enabling proactive content moderation.

Problem Description

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).

Dataset Source

The dataset can be accessed at the following link: https://www.kaggle.com/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv. It includes labelled Twitter data, with tweets classified as hate speech (label: 1) or non-hate speech (label: 0), specifically prepared for hate speech detection research

The dataset consists of two files:

Train Data which consists of **31,962** observations and **3 features**

Test Data which consists of **17,197** observations and **2** features

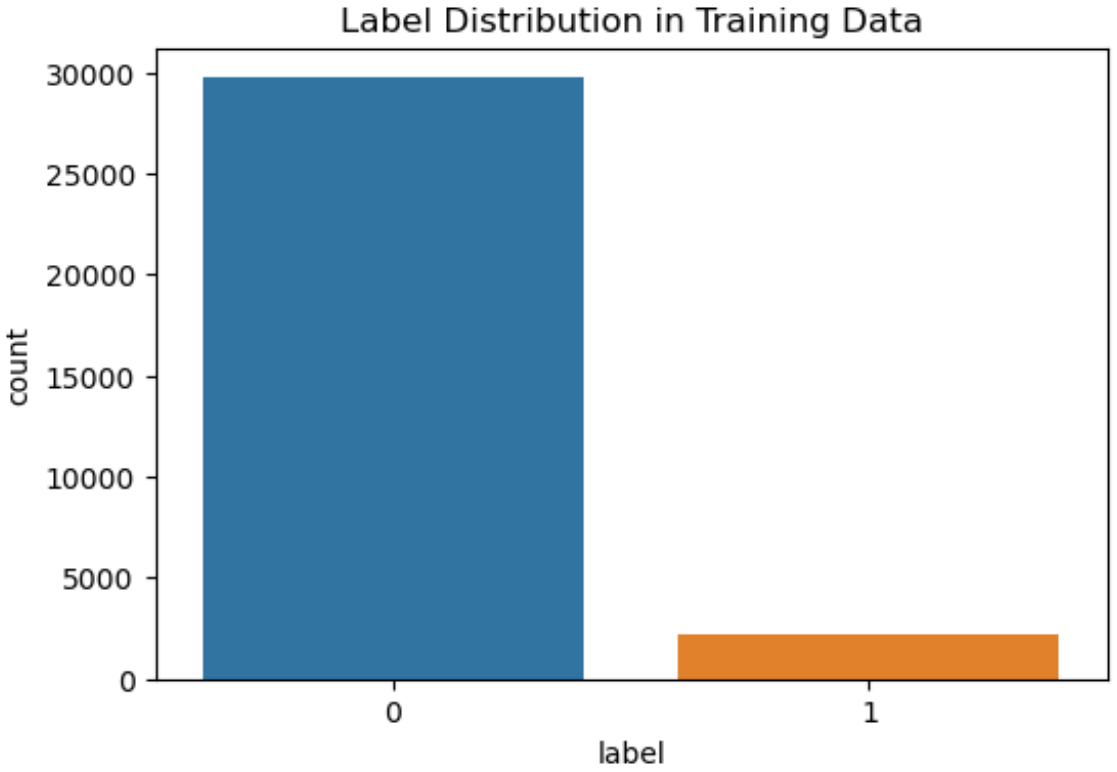
Data Pre-processing

TEXT CLEANING: Used regular expressions (regex) to clean the text by removing URLs, and special characters, and converting all text to lowercase. This ensured consistent input for the model and reduced noise in the data.

TOKENIZATION: Applied **BERT tokenizer** to split the cleaned text into tokens (small units) and convert them into numerical representations. This step enabled BERT to understand the context and meaning of each tweet effectively.

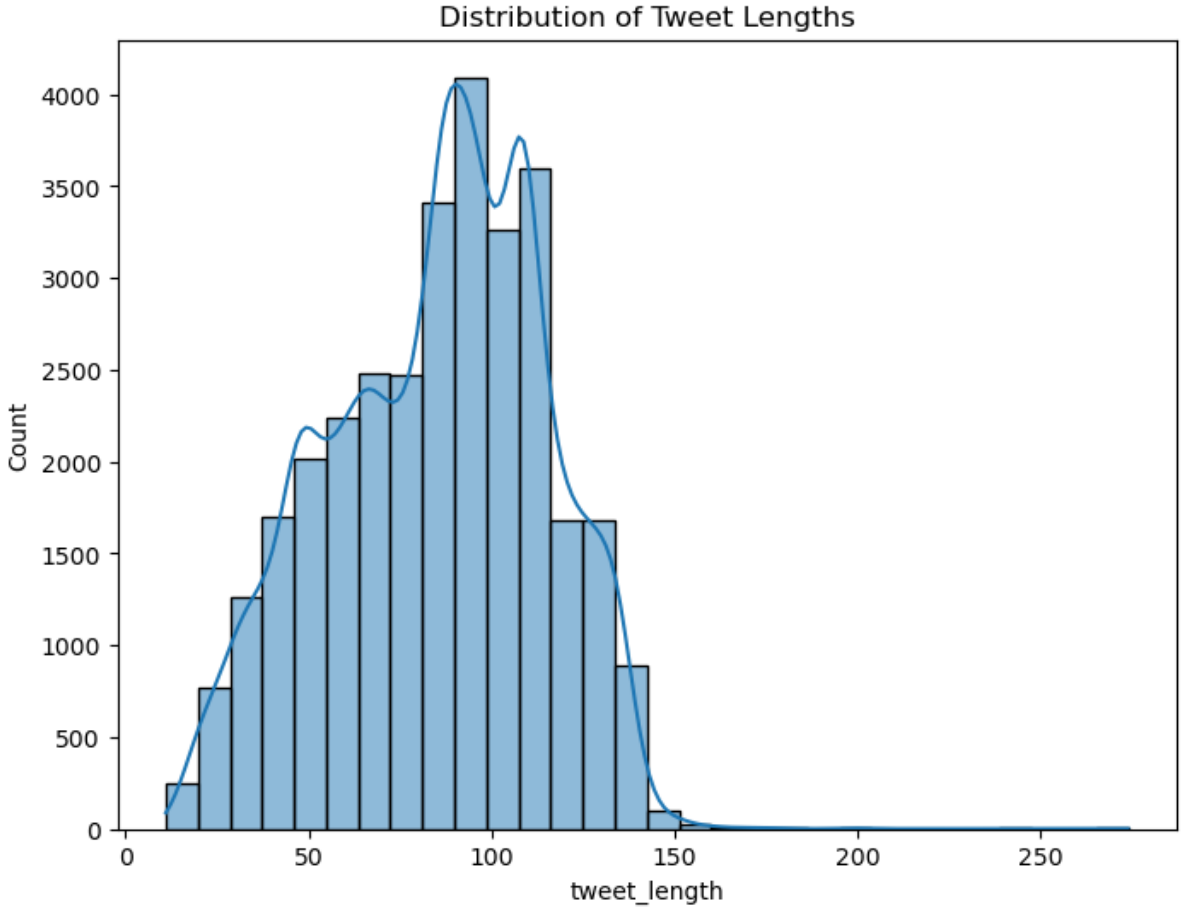
Exploratory Data Analysis

We initiated our analysis by assessing the distribution of labels and identified an imbalance in the dataset. To address this, We applied **class weights** in Transformers and Logistic Regression to emphasize the minority class and used **scale_pos_weight** in XGBoost to address the imbalance effectively



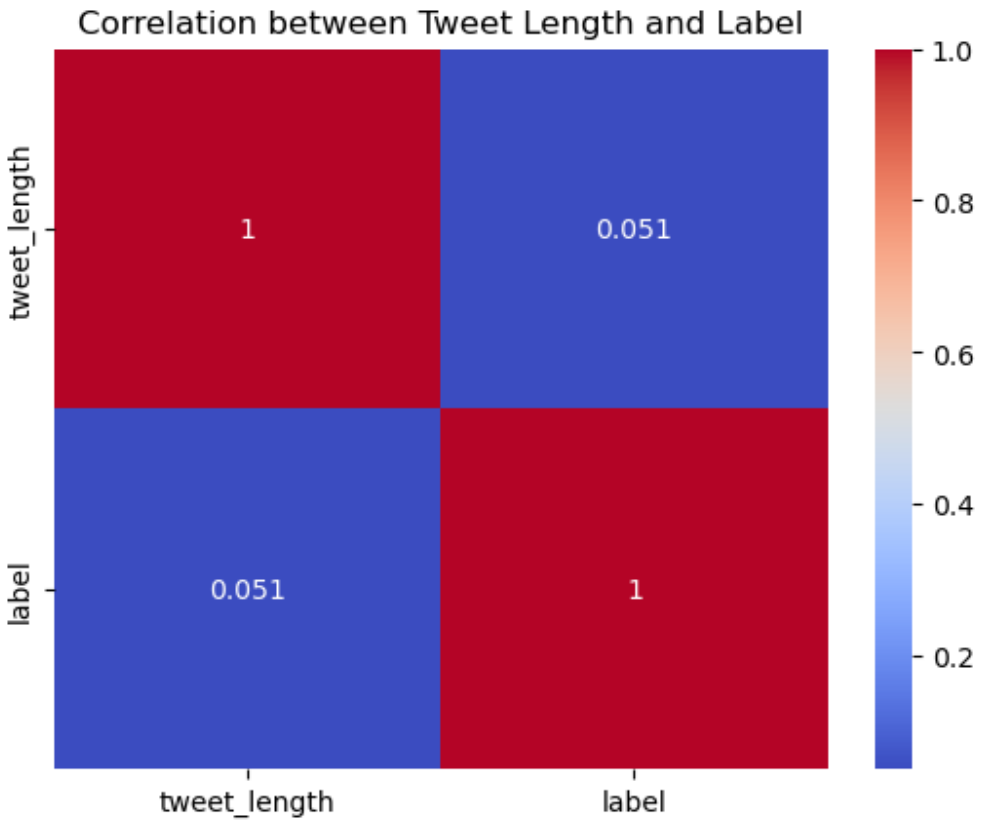
Exploratory Data Analysis

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

Model Selection

Chosen Models:

We used **Transformers (BERT)** as the primary model for classification due to its strong language understanding. **Logistic Regression** and **XGBoost** were included for comparison to assess model effectiveness.

Feature Representation

TF-IDF vectorization was applied for Logistic Regression and XGBoost, while the **BERT tokenizer** transformed text data into context-aware embeddings for the Transformer model.

Model Training and Evaluation Matrics

Training and Testing Setup:

Data was split into **80%** training and **20%** testing sets to ensure fair evaluation.

Evaluation Metrics:

We measured **accuracy, precision, recall,** and **F1-score** to assess each model. F1-score and recall were prioritized for capturing hate speech accurately.

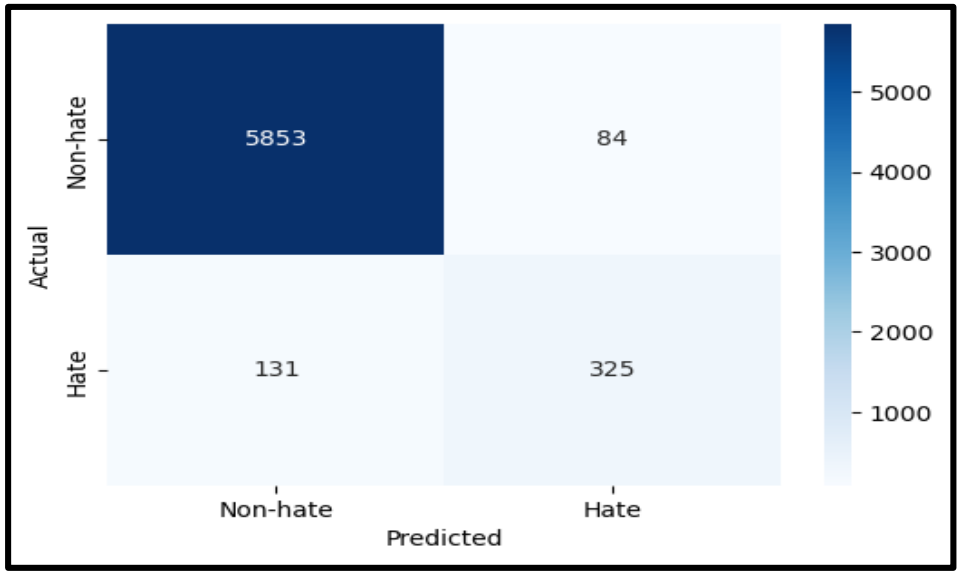
Transformer Model

Performance Summary: The Transformer model achieved high accuracy **(97%)** with strong F1-scores across both classes, particularly excelling in non-hate detection.

Confusion Matrix: Showed balanced performance and low misclassification across classes.

Insights: This model provided high reliability and generalizability, performing well on both hate and non-hate speech.

	precision	recall	f1-score	support
Non-hate	0.98	0.99	0.98	5937
Hate	0.79	0.71	0.75	456
accuracy			0.97	6393
macro avg	0.89	0.85	0.87	6393
weighted avg	0.97	0.97	0.97	6393



Logistic Regression Model

Performance Summary: Logistic Regression achieved high recall (**84%**) for hate speech, catching most hate tweets but with lower precision.

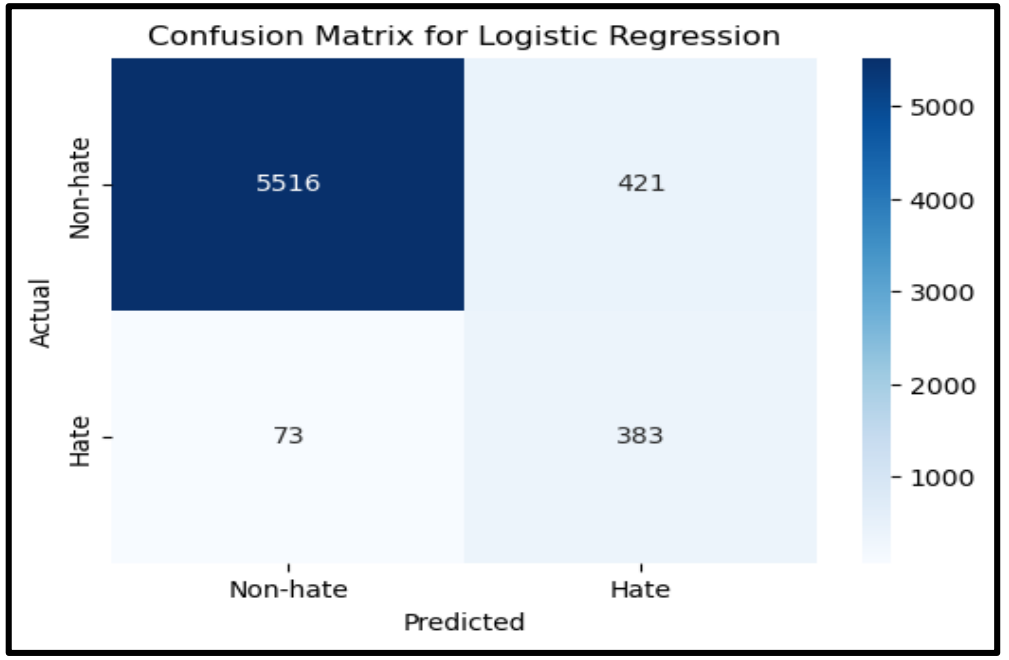
Confusion Matrix: Revealed high true positives for hate but also more false positives.

Insights: The model is sensitive to hate speech but tends to mislabel some non-hate tweets as hate.

Logistic Regression Results:

	precision	recall	f1-score	support
0	0.99	0.93	0.96	5937
1	0.48	0.84	0.61	456
accuracy			0.92	6393
macro avg	0.73	0.88	0.78	6393
weighted avg	0.95	0.92	0.93	6393

Accuracy: 0.9227279837322071



XGBoost Model

Performance Summary: XGBoost performed well on non-hate speech with high accuracy but struggled with lower recall and precision for hate speech..

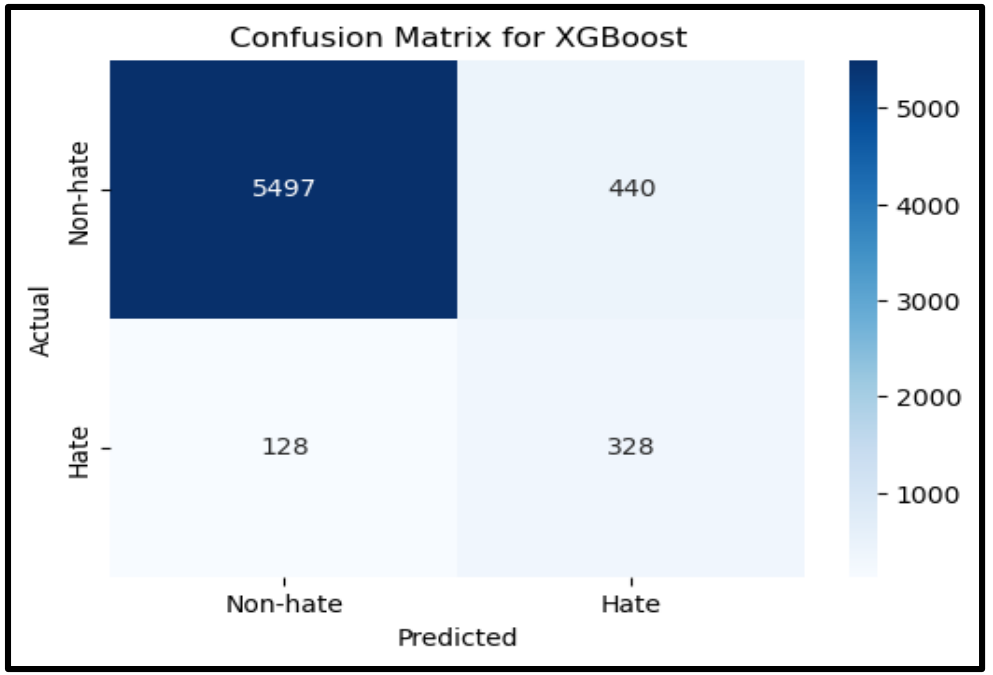
Confusion Matrix: Showed lower accuracy in distinguishing hate tweets, with more false negatives..

Insights: XGBoost performed reliably for non-hate tweets but had limited effectiveness in hate speech detection due to lower sensitivity..

XGBoost Results:

	precision	recall	f1-score	support
0	0.98	0.93	0.95	5937
1	0.43	0.72	0.54	456
accuracy			0.91	6393
macro avg	0.70	0.82	0.74	6393
weighted avg	0.94	0.91	0.92	6393

Accuracy: 0.9111528234005944

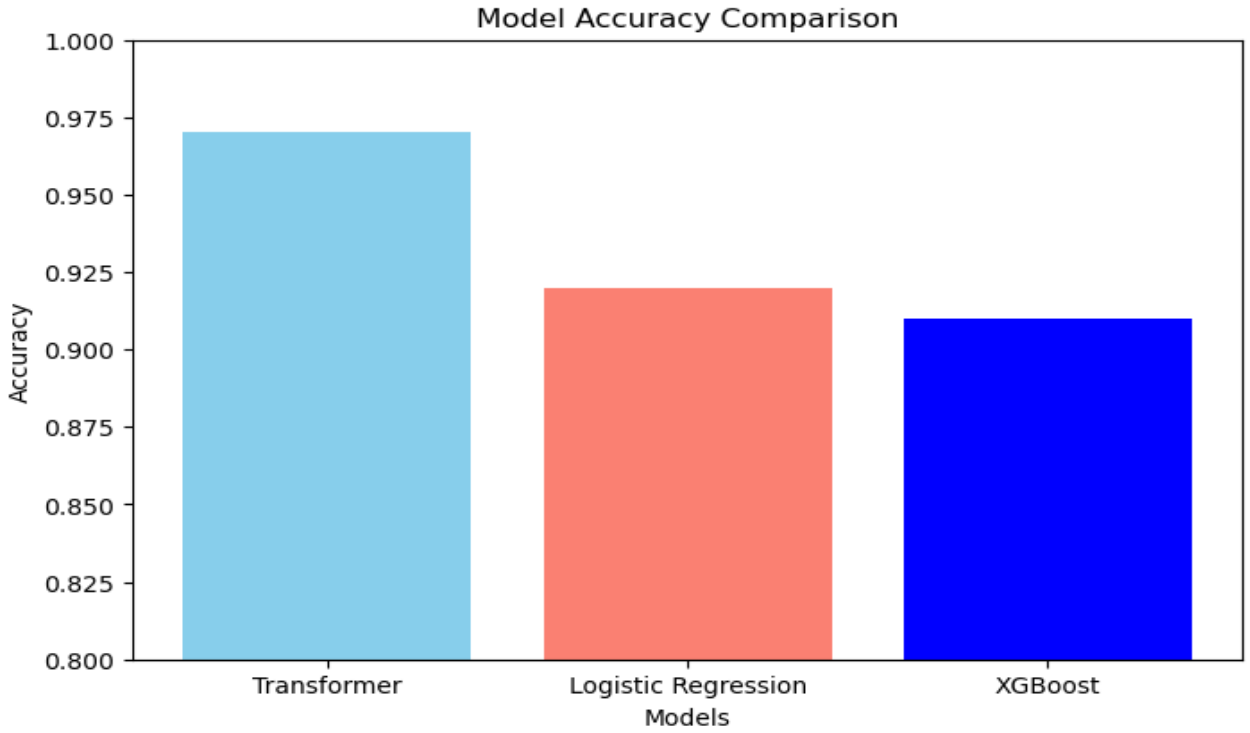


Comparison of Model Performance

Models were compared based on accuracy, F1-score, precision, and recall. The Transformer model emerged as the best performer across all metrics.

Transformers provided the most balanced and accurate results for both classes, making it the optimal choice for this project.

Model	Accuracy
Transformer	**97%**
Logistic Regression	**92%**
XGBoost	**91%**



Recommendation

Based on overall performance, **Transformers (BERT)** is recommended for deployment due to its high accuracy and balanced detection across both classes.

Further improvements could include **data augmentation**, **fine-tuning**, and **ensemble approaches** to enhance model robustness. Testing on larger datasets and implementing real-time monitoring could also improve practical applications.

CONCLUSION

This project demonstrated the effectiveness of Transformers for hate speech detection, achieving high accuracy and balanced performance.

Team Members

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