Judith Tesfaye • Springboard Data Science Career Track

Capstone 3 – AI vs Human Content Detection

08

2025

Agenda

01 AI writing tools are widely used in schools, hiring and media

02 It’s harder to tell AI-generated content from human writing

03 I explore a privacy-friendly approach using only structured metrics.

Goals & Success Metrics

- Goal: Classify AI vs human-written content without raw text.
- Constraint: Use only numeric readability/style metrics + content_type.
- Success: Accuracy ≥ 0.90 with balanced precision and recall.

Dataset Overview

•File: ai_human_content_detection_dataset.csv

•Features: word_count, lexical_diversity, sentence_length, grammar_errors, sentiment_score, etc.

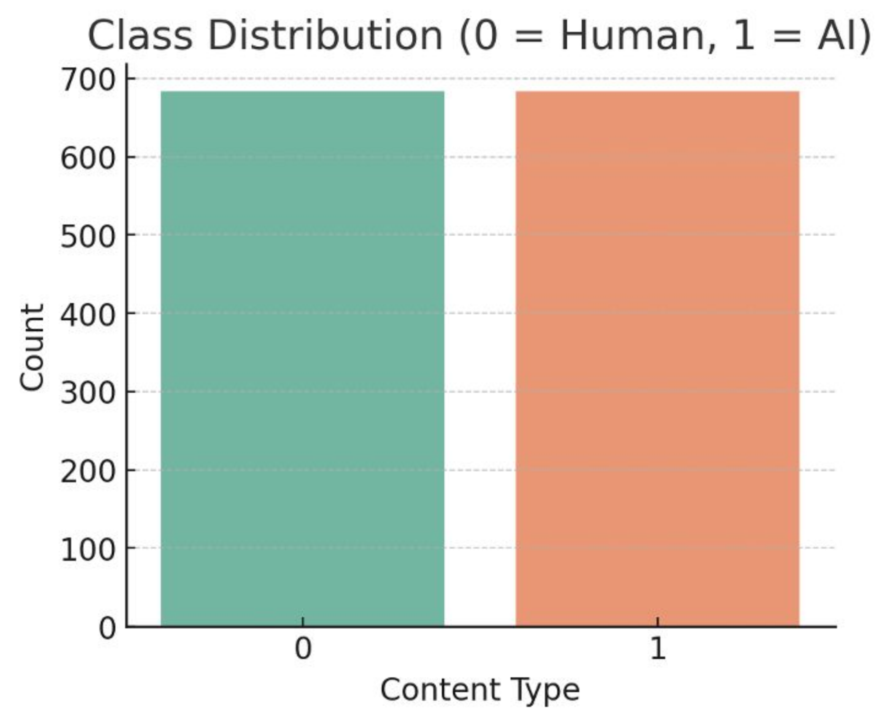
•Target: label (0 = Human, 1 = AI) • Categorical: content_type

Data Wrangling & Preparation

- Dropped raw text; kept structured metrics only.
- Scaled numeric features; one-hot encoded content_type.
- Train/Test split (80/20) with stratification.

EDA Visual: Class Distribution

The dataset is balanced across human (0) and AI (1) classes.
Balanced classes help prevent biased models and misleading metrics.
This supports fair training, validation, and model comparison.

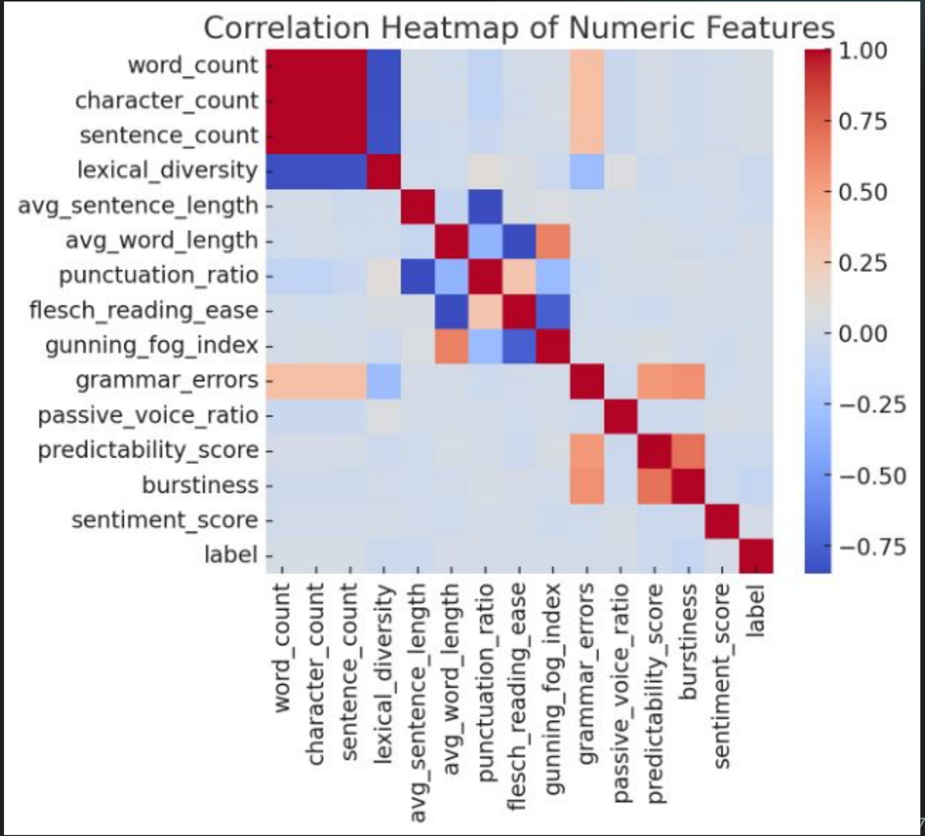


EDA Visual: Correlation Heatmap

Correlations highlight relationships among numeric features.

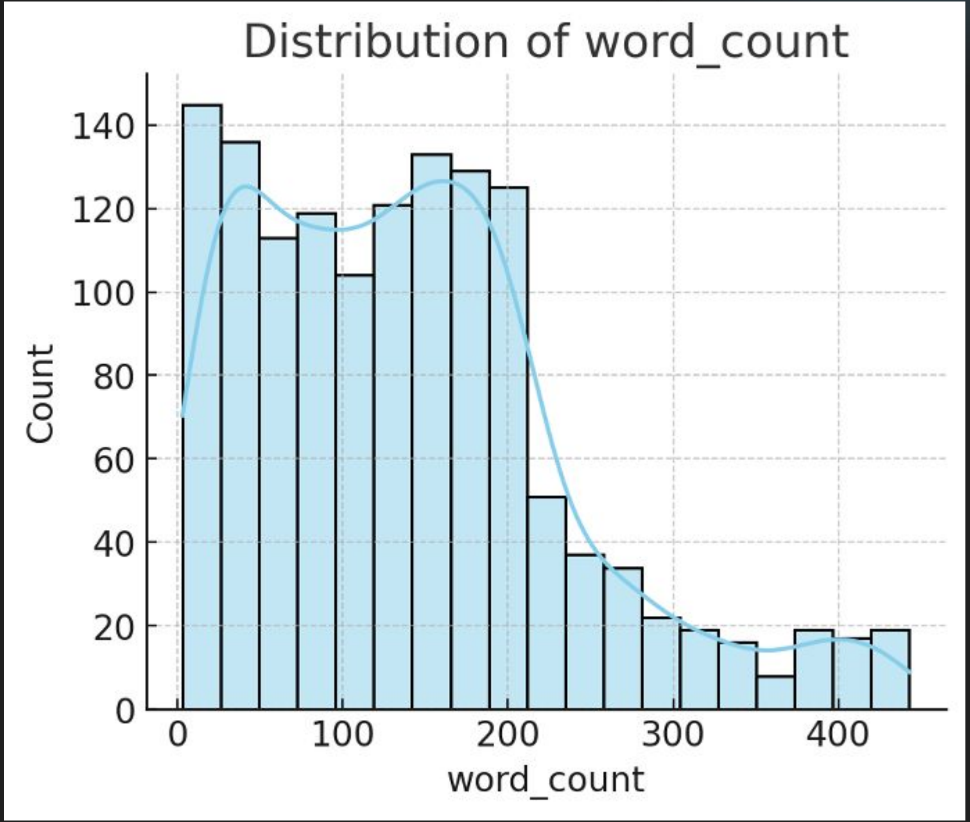
Highly related features may be redundant; weak ones can still add signal.

This view helps explain model behavior and feature selection.



EDA Visual: Feature Distribution

Correlations highlight relationships among numeric features.
Highly related features may be redundant; weak ones can still add signal.
This view helps explain model behavior and feature selection.



Modeling Approach

Logistic Regression

interpretable baseline.

Random Forest

handles non-linearities and interactions.

XGBoost

boosted trees for best accuracy and robustness.

Key Findings

- XGBoost achieved the highest accuracy among tested models.
- Random Forest also performed strongly on structured features.
- Structured metrics alone can effectively detect AI-generated content.

Recommendations & Applications

- Use XGBoost in a lightweight service for screening content.
- Integrate into plagiarism checks, moderation tools, and resume filters.
- Monitor drift and retrain with updated datasets periodically.

Conclusion & Next Steps

cOnclusion

- We can detect AI vs human content without using raw text.
- Next: add more features, tune hyperparameters, and expand data.
- Consider fairness checks across different content types.

Thank you!