Company Name Quarter Date

Final Project Presentation

By: Rucha Deshpande

Introduction

Research Question: Does being a native English speaker affect the likelihood of an AI detector making classification mistakes?

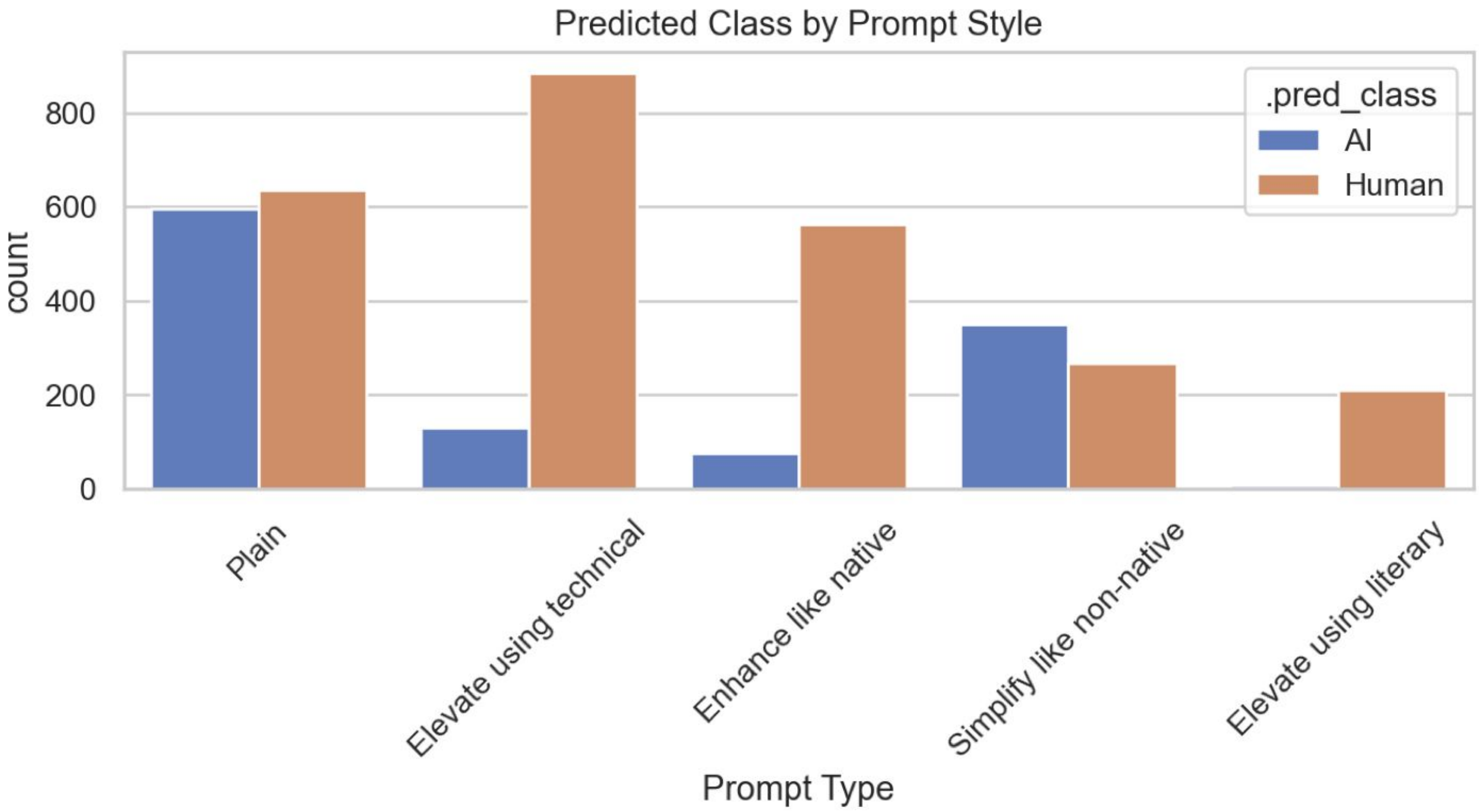
Motivations: I chose this research question and dataset because of the prevalence of language bias and AI detector tools used in academia today. Misclassifications have the power to ruin careers and lives, so I found this topic extremely relevant especially considering the rampant use of AI in our lives today.

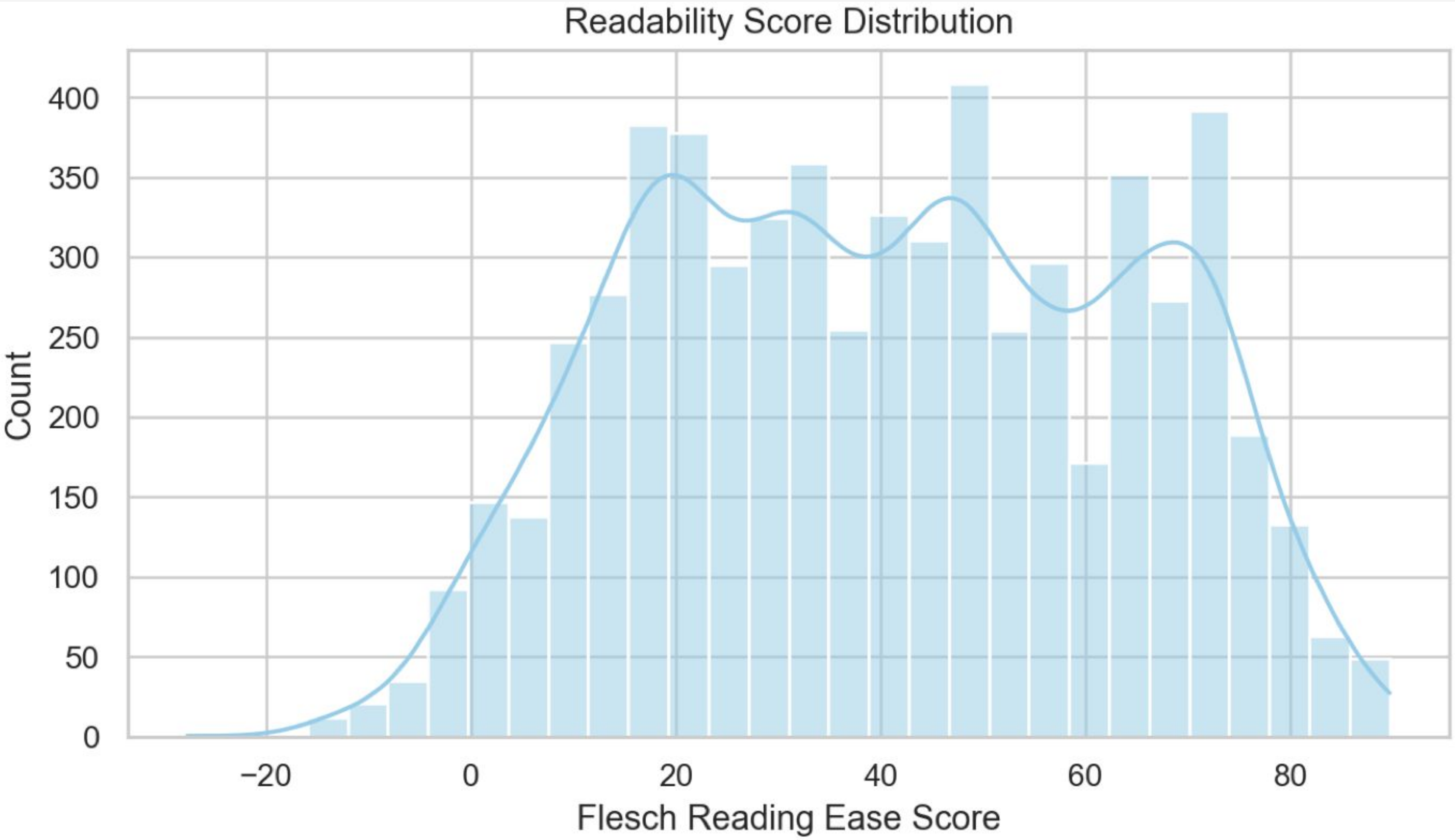
Data Wrangling

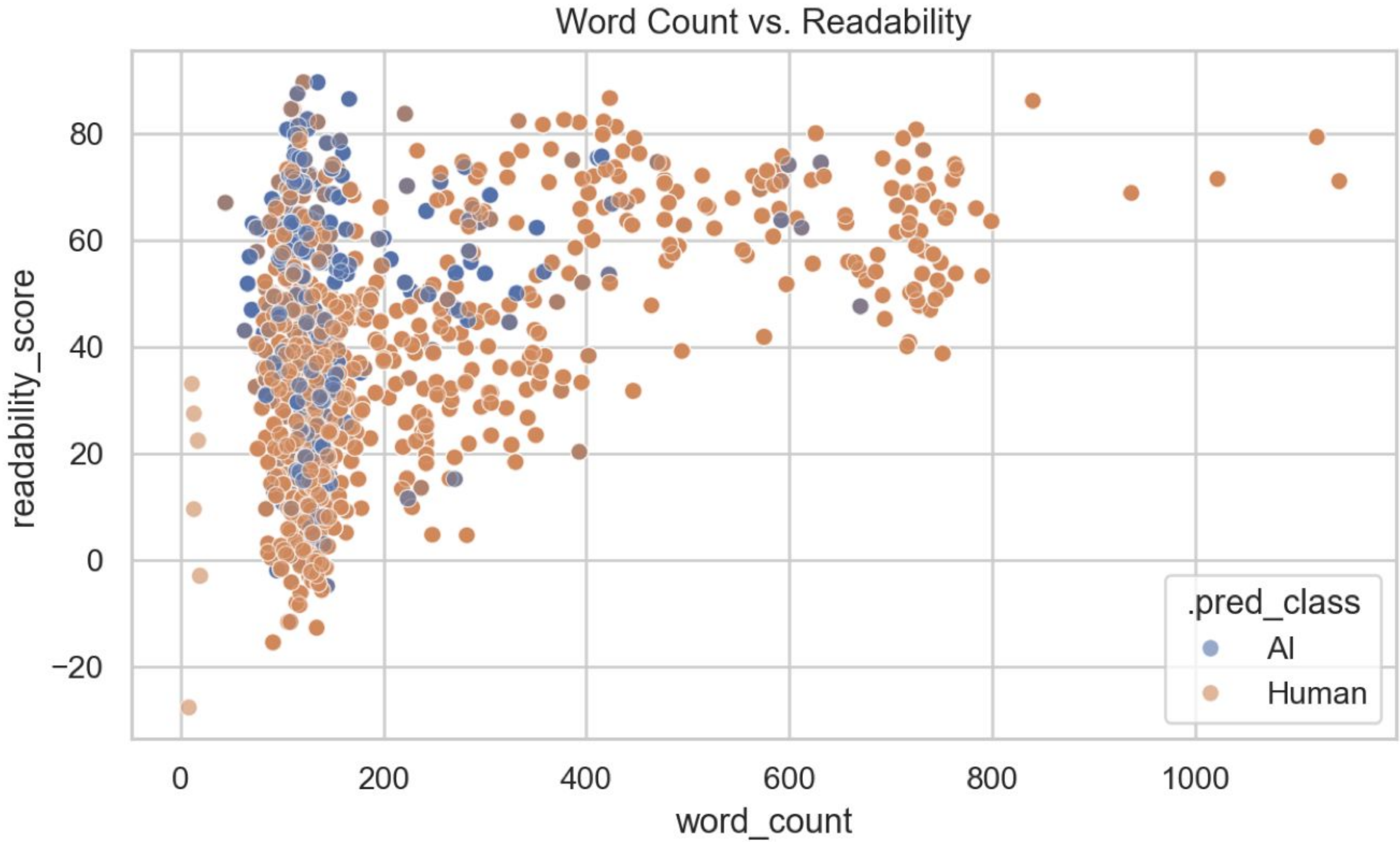
- The dataset I worked with was Simon Couch’s detectors R package
- Challenges:
    - .json -> .csv format
    - Preprocessing was difficult because for each document included in the raw dataset, I had to generate word/character/sentence count, average word length, and unique word count
        - Also extracted readability score (FLESCH), sentiment (VADER), keywords, and parts of speech counts
    - Main difficulty arose when trying to combine all of these variables into a single, enriched dataset without missing values
- Tools used: Python 3.13
    - pandas
    - numpy
    - seaborn/matplotlib for graphs
    - scikit-learn for modeling (train/test split)

EDA

- sentiment_score (VADER)
  - Most documents clustered around neutral
- readability_score (FLESCH)
  - Fairly low scores which suggests that essays uploaded are complex
- Native vs Non-native:
  - Native written texts had higher readability scores and longer average word length.
  - Some detectors more likely to classify non-native texts as AI (bias)
  - Non-native texts were usually simpler and more direct
- Prompts and AI-generated text
  - "Simplify like a non-native" - Lower complexity, higher misclassification as human
  - “Elevate using literary" - More adjectives/adverbs, often detected as AI
  - "Enhance like native" - Balanced structure but sometimes still flagged


Predicted Class by Prompt Style
.pred_class
AI
Human
count
0
200
400
600
800
Plain
Elevate using technical
Enhance like native
Simplify like non-native
Elevate using literary
Prompt Type


Readability Score Distribution
Count
0
50
100
150
200
250
300
350
400
−20
0
20
40
60
80
Flesch Reading Ease Score


Word Count vs. Readability
readability_score
80
60
40
20
0
−20
0
200
400
600
800
1000
word_count
.pred_class
● AI
● Human

Model Overview

Objective: Predict whether a text is AI-generated or human-written based on linguistic features extracted during EDA.

Model Used: Logistic Regression

Features Included: text length (word & sentence count), readability scores, sentiment scores, part of speech counts, keyword presence indicator, prompt type, and native/nonnative classification

Monte-Carlo

Objective: To assess variability of being a native english speaker on the likelihood of misclassification

Why use a Monte Carlo Simulation? Because the dataset has limited samples, and we want to understand the distribution of the logistic regression coefficient (β_1) under repeated sampling of the native variable.

Monte-Carlo Design

Simulation Design:

1. Start with logistic regression on the original data predicting misclassification (response variable) using native (predicting variable).
2. Extract coefficients (intercept β_0 and slope β_1)
3. Run 1000 simulations:
4. Resample native variable with replacement (bootstrapping).
5. Generate simulated binary response using logistic model with original coefficients.
6. Fit logistic regression on simulated data and record β_1 estimate.

Monte-Carlo Findings

Summary of Main Findings

- The distribution of simulated β_1 estimates is centered closely around the original logistic regression estimate, showing that the effect of native English on misclassification is stable
- The original estimate lies near the peak of the distribution, confirming the reliability of our initial model.
- Interpretation: Being a non-native English speaker significantly influences the probability of AI misclassification. This effect is not likely due to random variation in the data.
- Monte Carlo simulation adds confidence by showing the coefficient’s variability and helps validate the logistic regression’s conclusions

Observations

Research Question: How accurately can AI models detect AI-generated vs. human-written essays, and what factors affect misclassification rates?

Final Insights:

- AI misclassifies texts differently based on speaker nativeness and writing prompts.
- Native English speakers’ texts are less likely to be misclassified.
- Prompt style also influences detection accuracy, indicating AI models are sensitive to linguistic nuances.
- This highlights both strengths and limitations in current AI detection systems.