GPT Detector Efficacy

rhymond.1, denofsky.2, kleinholz.3, ling.273

Introduction

- GPT detectors are used to distinguish human-written vs. AI-generated text.
- Recent studies show they may be biased against non-native English writers (Liang, W. et al 2023)
- Such misclassifications can have serious implications, including unfair academic penalties and misjudged credibility in professional work.
- Our project investigates the extent and nature of these biases across different GPT detectors.

The Objective

We want to examine whether GPT detectors disproportionately misclassify writing from non-native English speakers as AI-generated, exploring the overall accuracy of each model.

Our goal: understand not only if such biases exist, but also how they appear through different types of errors and whether certain detectors are more prone to misclassification.

Data Overview

- **Source**: TidyTuesday (July 18, 2023), based on Liang et al. (2023).
- Key Variables:
  - `kind`: Ground truth (“Human” or “AI”)
  - `.pred_class`: Detector’s prediction
  - `native`: Whether writer is a native English speaker
  - `detector`: Detector used
  - `.pred_AI`: Probability assigned to being AI-generated
  - n = 6185

Data Preparation

- Added `correct` binary field; determines if predictions are correct
- Added `type1error` binary field; determines if outcomes are false positives (Type I errors)

Methods

We use binomial logistic regression to model the correctness of each AI detector. This should allow us to uncover model biases. For our regression we focused on only the essays written by real humans to better understand when the detectors were incorrect. We ended up with n = 2468.

Predictors:

- Detector used (`Detector_j`)
- Native English status (`Native`)
- Interaction between `Detector_j` and `Native`

Model Format

Logistic Regression for Model Correctness

Y_i (classification correctness) for $i = 1, \ldots, n$ are mutually-independent Bernoulli variables with binary responses.

$$\text{logit}\big(P(Y_i = 1)\big) = \beta_0 + \sum_{j=1}^{J} \beta_j \cdot \text{detector}_j + \beta_N \cdot \text{native} + \sum_{j=1}^{J} \beta_{jN} \left(\text{detector}_j \cdot \text{native}\right)$$

Results: Misclassification Metrics

Are non-native writers more likely to be misclassified?

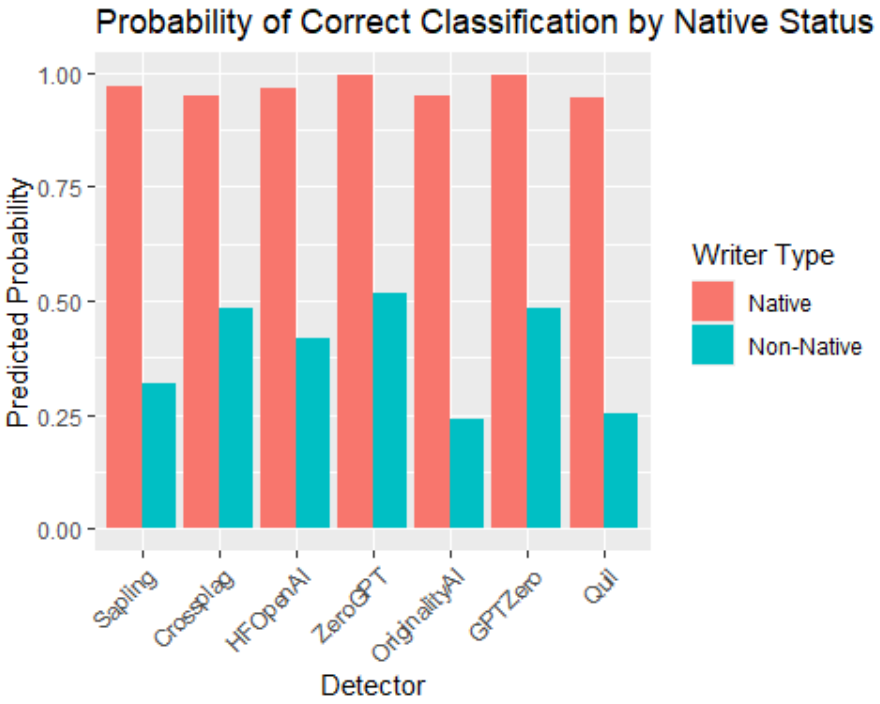
$\beta_{native=YES}$ is one of the covariates from the GLM with value 2.953, indicating that native speakers are more likely to be correctly classified compared to non-native speakers because it is greater than 1.

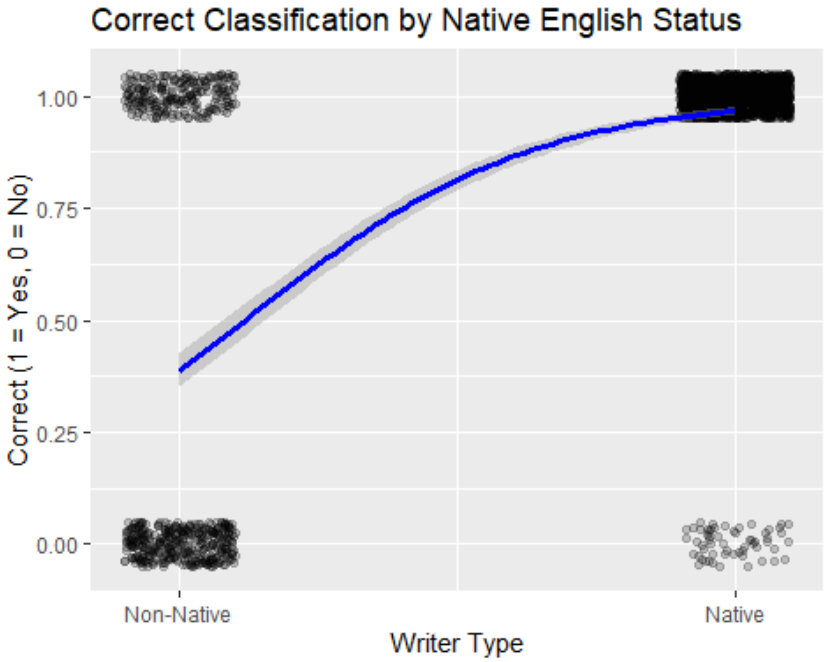
We can use exp() to get the odds ratio:

```
## [1] 19.16336
```

The odds of a native speaker being correctly classified is 19x greater than the odds of a non-native speaker being correctly classified.

Analysis


Probability of Correct Classification by Native Status
Predicted Probability
1.00
0.75
0.50
0.25
0.00
Sapling
Crossplag
HFOpenAI
ZeroGPT
OriginalityAI
GPTZero
Quil
Detector
Writer Type
Native
Non-Native


Correct Classification by Native English Status
Correct (1 = Yes, 0 = No)
1.00
0.75
0.50
0.25
0.00
Non-Native
Native
Writer Type

Results: Detector Bias

Are some detectors more biased?

```
## # A tibble: 6 × 4
##   term                                estimate p_value odds_ratio
##   <chr>                                  <dbl>   <dbl>      <dbl>
## 1 detectorGPTZero:nativeYes              2.13  0.00869       8.39
## 2 detectorHFOpenAI:nativeYes             0.757 0.136         2.13
## 3 detectorOriginalityAI:nativeYes        1.15  0.0194        3.14
## 4 detectorQuil:nativeYes                 0.938 0.0795        2.55
## 5 detectorSapling:nativeYes              1.23  0.0427        3.41
## 6 detectorZeroGPT:nativeYes              2.00  0.0138        7.35
```

Discussion

- Clear evidence of bias in AI detectors against non-native writers.
- False positives (humans flagged as AI) are prevalent for non-native speakers.
- Suggests caution in using GPT detectors for grading or evaluation purposes.

Conclusion

- GPT detectors generally classify native writers correctly, as humans.
- All GPT detectors exhibit some level of bias against non-native writers, disproportionately misclassifying them as AI.
- GPT detectors vary in their degree of bias against non-native writers. All of them are biased, but some to a lesser degree.
- Future work should focus on fairness-aware detection algorithms.