Evaluating GPT Detector Performance

Accuracy and Bias in Human Text Classification

Catherine Ling, Jonathan Rhymond, Jake Denofsky, Meia Kleinholz

April 23, 2025

Abstract

GPT detectors are increasingly used in academic and professional settings to identify AI-generated text. However, recent research raises concerns about their fairness, particularly towards non-native English speakers. In this report, we analyze publicly available data from the TidyTuesday project, which builds on work by Liang et al. (2023), to assess the presence and extent of bias in detector outputs. Using logistic regression on over 2400 human-written samples, we quantify the likelihood of false positives and correct classification by native status. Our findings show a significant disparity: native speakers are approximately 19 times more likely to be correctly identified as human compared to non-native speakers. These results suggest the need for more equitable detection systems in real-world applications.

Introduction

As generative AI tools become increasingly integrated into academic and professional environments, there is a growing reliance on GPT detectors — machine learning models trained to distinguish between human-written and AI-generated text. These tools are often used to enforce academic integrity policies or assess the authenticity of written content. However, emerging research suggests that these detectors may not be entirely impartial. For instance, Liang et al. (2023) demonstrate that GPT detectors can disproportionately flag text written by non-native English speakers as AI-generated, even when it is entirely human-written. This raises critical concerns about fairness, especially in educational and hiring contexts where such misclassifications could lead to unwarranted academic penalties or diminished professional credibility.

In this project, we aim to systematically evaluate the presence and degree of bias in GPT detectors. Specifically, we apply statistical methods and logistic regression models to analyze the classification outputs of several widely used detectors. By quantifying disparities in misclassification rates across different groups of writers, we seek to understand the underlying factors contributing to biased predictions and assess the robustness and fairness of these detection tools.

Question of Interest

We aim to investigate whether GPT detectors disproportionately misclassify writing by non-native English speakers as AI-generated. Specifically, we will examine overall accuracy and the frequency of false positives (Type I errors).

Our objective is not only to determine the existence of such biases, but also to analyze how they manifest through various types of errors. Additionally, we seek to identify whether certain detectors are more susceptible to these misclassifications.

Data Overview

- **Source**: Data obtained from the TidyTuesday project (July 18, 2023), based on the dataset published by Liang et al. (2023).
- Key Variables:
  - `kind`: Ground truth label indicating whether the text was written by a human or generated by AI.
  - `.pred_class`: The classification output from the GPT detector (AI or Human).
  - `native`: Indicates whether the author is a native English speaker.
  - `detector`: The specific GPT detection model used for evaluation.
  - `.pred_AI`: The probability score assigned by the detector to the text being AI-generated.
  - $n = 6185$

Data Preparation

kind	pred_ai	pred_class	detector	native	name	model	document_id	prompt	correct
Human	0.9999942	AI	Sapling	No	Real TOEFL	Human	497	NA	FALSE
Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human	278	NA	FALSE
Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human	294	NA	TRUE
Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human	855	NA	TRUE
Human	0.0000000	Human	ZeroGPT	Yes	Real CS224N	Human	781	NA	TRUE
Human	0.9999993	AI	Sapling	No	Real TOEFL	Human	460	NA	FALSE

- Added `correct` binary field: determines if predictions are correct.

Methods

To evaluate the accuracy and potential biases of different AI detectors, we employed a binomial logistic regression model. This statistical method is well-suited for our analysis because the outcome variable (whether the detector's classification was correct) is binary. Logistic regression will allow us to estimate the probability of a correct classification while accounting for multiple categorical predictors and their interactions.

Our analysis focused exclusively on essays written by real human authors ($n = 2468$). By isolating human-written texts, we aimed to better understand the conditions under which detectors falsely label human work as AI-generated, i.e., false positives (Type I errors).

To investigate potential biases in AI-generated text detectors, we conducted statistical analysis using a binomial logistic regression model. This approach is appropriate for our study because the dependent variable (whether a detector correctly classifies a human-written essay) is binary. Logistic regression enables us to model the probability of a correct classification as a function of multiple categorical predictors and their interaction effects.

Our dataset comprises 2,468 essays, all written by verified human authors. We intentionally excluded AI-generated texts from the analysis to focus on false positives — cases where human-written work is misclassified as AI-generated (i.e., Type I errors). This allowed us to directly assess whether certain detectors are systematically biased against particular groups of writers.

Model Specification

Let Y_i denote the binary outcome for the i^{th} observation, where $Y_i = 1$ indicates that the detector's classification was correct, and $Y_i = 0$ otherwise. We assume the Y_i follow independent Bernoulli distributions and model the log-odds of a correct classification using the following logistic regression formulation:

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

Where: - β_0 is the intercept term - detector_j are binary indicator variables for each AI detector used (with one omitted as a baseline) - native_i indicates whether the essay author is a native English speaker - $\text{detector}_j \cdot \text{native}_i$ represents the interaction effect between detector type and native speaker status

This structure allows us to test whether the misclassification rates vary not only by detector and author background individually, but also whether some detectors are more biased than others against non-native speakers.

Predictor Variables:

- Detector (`detector`): A categorical variable denoting which GPT detector was applied to the text (e.g. GPTZero,HFOpenAI, OriginalityAI)
- Native English Status (`native`): A binary variable indicating whether the essay author is a native English speaker.
- Interaction Term (`detector × native`): Captures potential detector-specific disparities in misclassification between native and non-native writers.

Model Outputs

We fit the model using a generalized linear model (GLM) with a logit link function. Coefficient estimates are provided in both log-odds form and as odds ratios for ease of interpretation. Odds ratios greater than 1 indicate a higher likelihood of correct classification, while values less than 1 suggest an increased probability of misclassification.

term	estimate	std.error	statistic	p.value	odds_ratio
(Intercept)	-0.066	0.210	-0.314	0.753	0.936
detectorGPTZero	0.000	0.297	0.000	1.000	1.000
detectorHFOpenAI	-0.267	0.299	-0.893	0.372	0.766
detectorOriginalityAI	-1.077	0.322	-3.341	0.001	0.341
detectorQuil	-1.018	0.320	-3.185	0.001	0.361
detectorSapling	-0.694	0.308	-2.256	0.024	0.500
detectorZeroGPT	0.132	0.297	0.445	0.657	1.141
nativeYes	2.953	0.332	8.904	0.000	19.161
detectorGPTZero:nativeYes	2.127	0.811	2.624	0.009	8.390
detectorHFOpenAI:nativeYes	0.757	0.509	1.490	0.136	2.133
detectorOriginalityAI:nativeYes	1.145	0.490	2.337	0.019	3.143
detectorQuil:nativeYes	0.938	0.535	1.754	0.079	2.555
detectorSapling:nativeYes	1.228	0.606	2.027	0.043	3.414
detectorZeroGPT:nativeYes	1.995	0.811	2.461	0.014	7.353

Results

1. Misclassification by Native English Status

Our regression analysis indicates a substantial disparity in classification accuracy based on native English status. The estimated coefficient for `native` is 2.953, corresponding to an odds ratio of $\exp(\beta_7 = 2.953) = 19.16$.

This means that, holding other factors constant, native English speakers are approximately 19 times more likely to be correctly classified as human compared to non-native speakers.

Visual Evidence

Figure 1: Probability of Correct Classification by Native Status

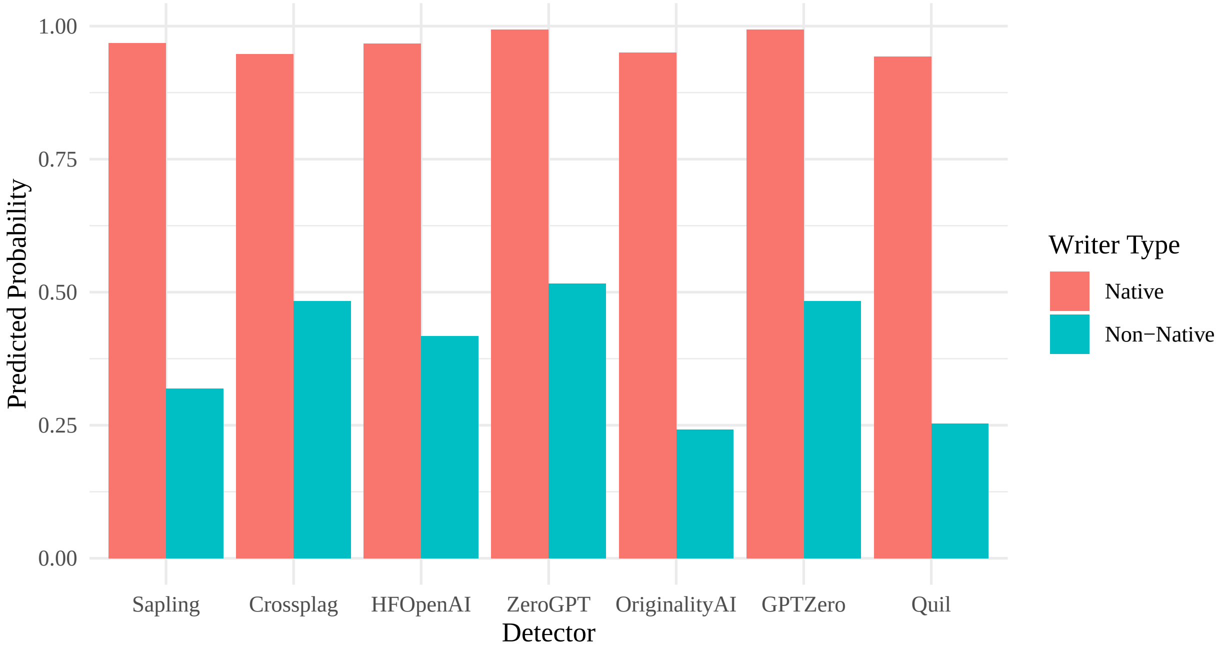

Figure 2: Correct Classification by Native English Status

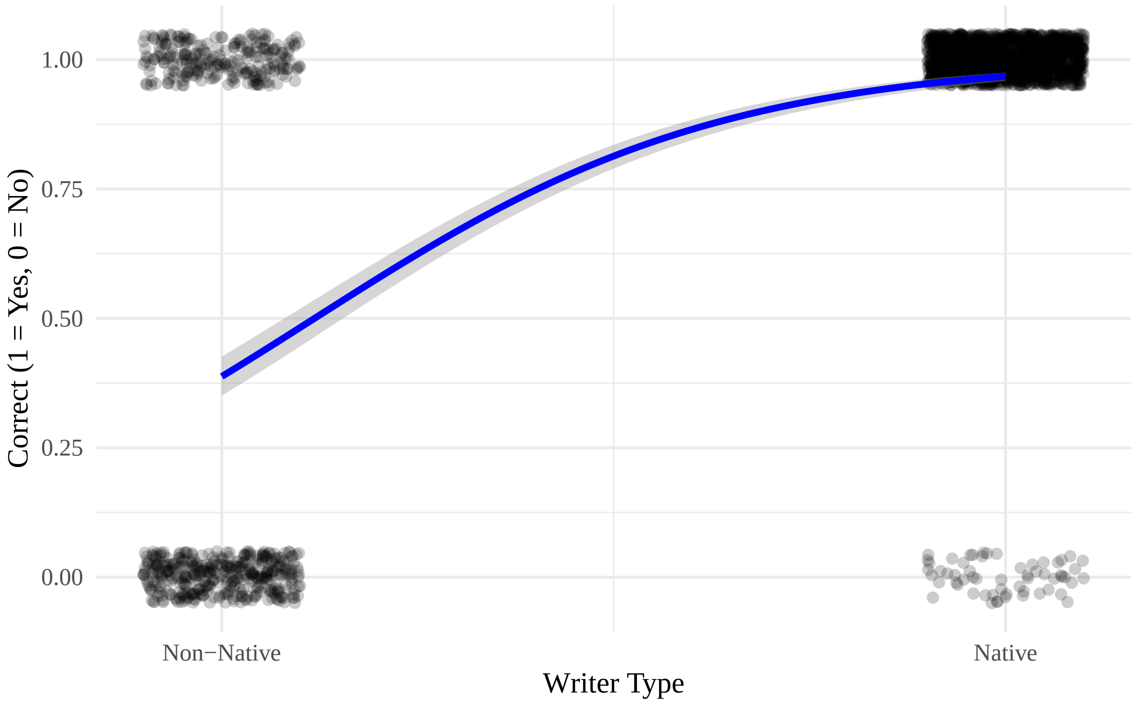

Figure 1 presents the predicted probabilities of correct classification across different detectors, disaggregated by native English status. Across all models, native speakers are consistently more likely to be correctly identified as human authors. This suggests a systemic bias against non-native writers across detection systems.

Figure 2 shows individual data points and the fitted logistic regression curve, illustrating the stark divide in classification correctness between native and non-native speakers. The visual reinforces the large effect size estimated by the regression model.

2. Detector-Specific Bias

To further examine how misclassification varies by detector, we analyzed the interaction terms between `detector` and `native` status. These interaction coefficients capture whether each detector shows disproportionately high false positive rates for non-native English writers.

term	estimate	p_value	odds_ratio
detectorGPTZero:nativeYes	2.1270695	0.0086926	8.390244
detectorHFOpenAI:nativeYes	0.7574418	0.1363492	2.132813
detectorOriginalityAI:nativeYes	1.1451229	0.0194335	3.142828
detectorQuil:nativeYes	0.9378838	0.0794718	2.554570
detectorSapling:nativeYes	1.2279871	0.0426975	3.414350
detectorZeroGPT:nativeYes	1.9951536	0.0138479	7.353332

As shown, GPTZero demonstrates the strongest bias, with non-native writers being over 8 times less likely to be correctly classified than their native counterparts. HFOpenAI, while still biased, exhibits a smaller disparity. Importantly, all detectors display statistically significant bias, as reflected by odds ratios above 1.

Discussion

Our findings confirm that current GPT detectors are not equally accurate across demographic groups, specifically disadvantaging non-native English writers. This bias is not uniform: certain detectors are more prone to misclassification than others, suggesting differences in training data, language modeling assumptions, or decision thresholds.

This has serious real-world implications. In academic settings, non-native students may face unjust accusations of AI misuse. In professional contexts, their credibility may be unfairly questioned. The presence of such bias calls into question the ethical deployment of these tools.

Limitations and Future Work

While informative, our analysis is subject to several limitations: - We focused exclusively on Type I errors (false positives). Future work should also examine Type II errors, where AI-generated content is incorrectly labeled as human. - The dataset includes only English-language essays. Further study is needed to assess whether similar biases appear in other languages or multilingual contexts. - We did not account for writing proficiency levels, which may confound the effect of nativeness. Future models could incorporate external proficiency assessments or grading rubrics.

Possible mitigation strategies include threshold calibration, adversarial training, or post-hoc bias correction using fairness-aware machine learning techniques.

Conclusion

Our investigation reveals that GPT detectors exhibit significant and systematic bias against non-native English speakers, leading to disproportionate false positive rates. This bias was observed across all detectors analyzed, though the magnitude varied by tool.

These results underscore the urgent need for bias-aware model development. As AI detection tools become more widely adopted in evaluative and high-stakes settings, addressing fairness concerns is not just a technical challenge, but an ethical imperative.