Non-native English essays disproportionately flagged as AI

Author: Michael Song

This notebook analyzes whether essays written by non-native English speakers are more likely to be falsely flagged as AI compared to native English speakers. Data for this analysis is taken from a 2023 edition of Tidy Tuesday, and is based off of a study done here

This study mainly serves as a way for me to learn some more modern R packages in the tidyverse, such as readr, dplyr, and ggplot2.

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(purrr)
```

```r
detectors <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/20
```

```
## ‘curl‘ package not installed, falling back to using ‘url()‘
```

We can see that each row in our data represents one essay and instance of a detector. We also see that our data has 10 unique categories for the name column, with "Real TOEFL" being the only one corresponding to non-native human-written essays.

```r
print(detectors)
```

```
## # A tibble: 6,185 x 9
##    kind  .pred_AI .pred_class detector    native name  model document_id prompt
##    <chr>    <dbl> <chr>       <chr>       <chr>  <chr> <chr>       <dbl> <chr>
##  1 Human 1.000    AI          Sapling     No     Real~ Human         497 <NA>
##  2 Human 0.828    AI          Crossplag   No     Real~ Human         278 <NA>
##  3 Human 0.000214 Human       Crossplag   Yes    Real~ Human         294 <NA>
##  4 AI    0        Human       ZeroGPT     <NA>   Fake~ GPT3          671 Plain
```

```
##   5 AI    0.00178  Human       Originality~ <NA>   Fake~ GPT4          717 Eleva~
##   6 Human 0.000178 Human       HFOpenAI     Yes    Real~ Human         855 <NA>
##   7 AI    0.992    AI          HFOpenAI     <NA>   Fake~ GPT3          533 Plain
##   8 AI    0.0226   Human       Crossplag    <NA>   Fake~ GPT4          484 Eleva~
##   9 Human 0        Human       ZeroGPT      Yes    Real~ Human         781 <NA>
##  10 Human 1.000    AI          Sapling      No     Real~ Human         460 <NA>
## # i 6,175 more rows
```

```r
print(detectors$name %>% unique() %>% length())
```

```
## [1] 10
```

```r
print(detectors$name %>% unique())
```

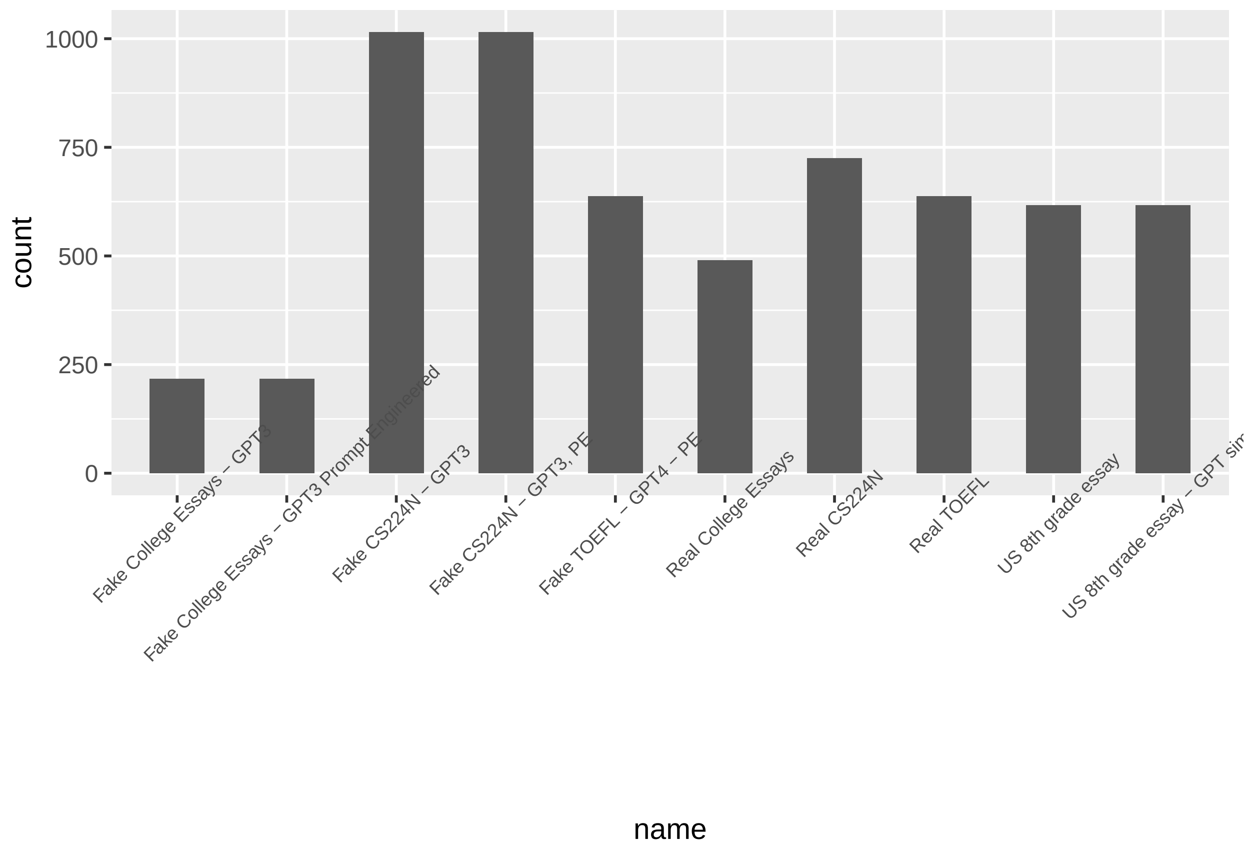
```
##  [1] "Real TOEFL"
##  [2] "Real College Essays"
##  [3] "Fake CS224N - GPT3"
##  [4] "Fake CS224N - GPT3, PE"
##  [5] "Real CS224N"
##  [6] "US 8th grade essay"
##  [7] "Fake TOEFL - GPT4 - PE"
##  [8] "Fake College Essays - GPT3"
##  [9] "US 8th grade essay - GPT simplify"
## [10] "Fake College Essays - GPT3 Prompt Engineered"
```

We want to create a separate column/flag to denote whether an essay is ai, human (native), or human (non-native) for future analysis. We also want to create a flag for whether the detector flagged the essay correctly or incorrectly.

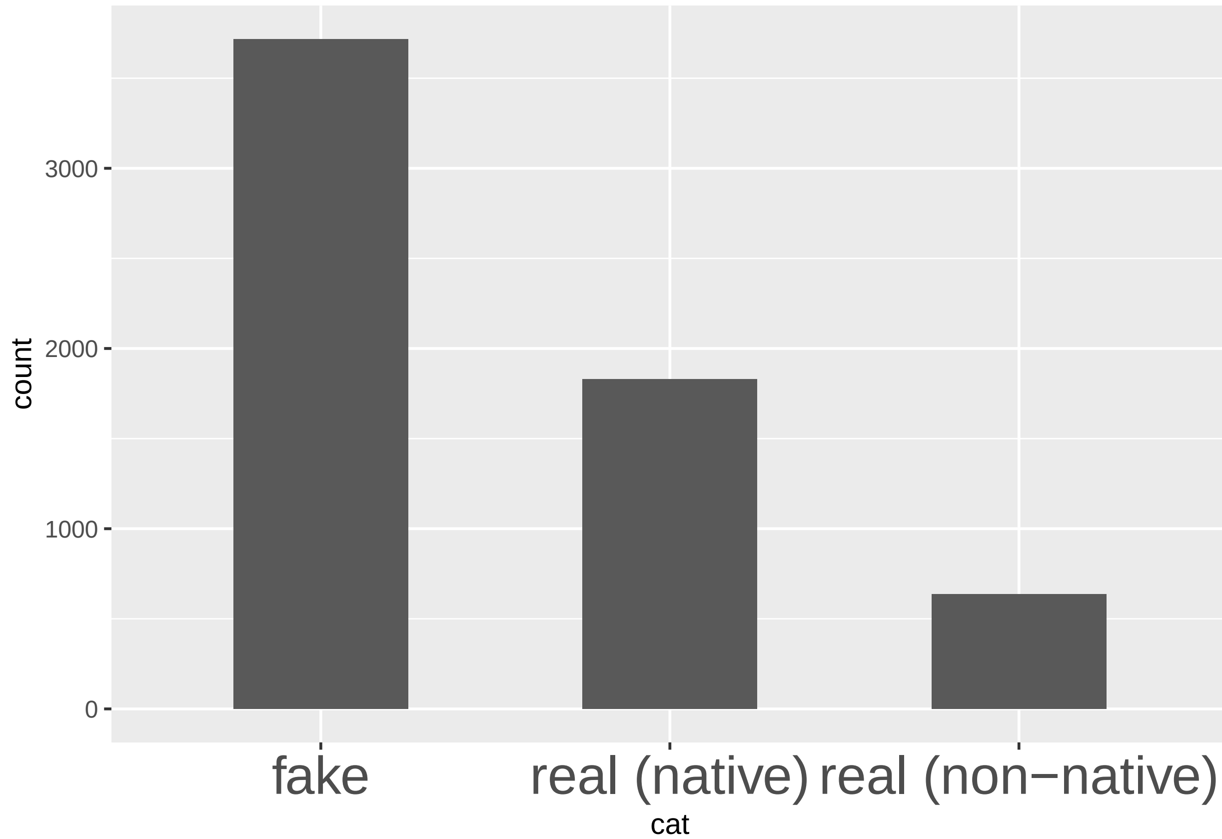
```r
detectors <- detectors %>%
  mutate(
    cat = case_when(
      kind == "AI" ~ "fake",
      native == "Yes" ~ "real (native)",
      TRUE ~ "real (non-native)"),
    right = case_when(kind == .pred_class ~ TRUE, TRUE ~ FALSE)
  )
```

We can visualize the original categories and our aggregate cat column to see the spread of observations.

```r
ggplot(detectors, aes(x = name)) +
  geom_bar(width=0.5) +
  theme(axis.text.x  = element_text(angle=45, size = 7))
```




```r
ggplot(detectors, aes(x = cat)) +
  geom_bar(width=0.5) +
  theme(axis.text.x  = element_text(size = 20))
```



We can see from the above charts that most of the observations come from AI-generated data.

Analysis

First, lets check the accuracy of our detectors without differentiating between native and non-native human essays

```
accuracy <- group_by(detectors, detector) %>% summarise(accuracy = sum(right) / n())
mean_accuracy <- mean(accuracy$accuracy)
print(accuracy)
```

```
## # A tibble: 7 x 2
##   detector      accuracy
##   <chr>            <dbl>
## 1 Crossplag        0.501
## 2 GPTZero          0.489
## 3 HFOpenAI         0.514
## 4 OriginalityAI    0.590
## 5 Quil             0.478
## 6 Sapling          0.5
## 7 ZeroGPT          0.517
```

```
print(mean_accuracy)
```

```
## [1] 0.5125621
```

Our mean accuracy is only ~51%, with accuracy not differing too wildly between different detectors. This is barely better than guessing.

Lets go ahead and compare the likelihood that a human (native) essay is flagged as AI compared to a human (non-native) essay

```
human.native <- filter(detectors, cat == "real (native)")
human.non.native <- filter(detectors, cat == "real (non-native)")
```

```
# getting fraction of native/non-native essays that are flagged as AI
native.flagged <- human.native %>% summarize(flagged = 1 - sum(right) / n())
non.native.flagged <- human.non.native %>% summarize(flagged = 1 - sum(right) / n())
print(native.flagged %>% pull)
```

```
## [1] 0.03222283
```

```
print(non.native.flagged %>% pull)
```

```
## [1] 0.6122449
```

We can see that the means seem to differ.

To formally test whether these proportions significantly differ, we can do a two-proportion z test

```
test.results <- prop.test(x = c(sum(!human.native$right), sum(!human.non.native$right)), n = c(nrow(huma
print(test.results)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(!human.native$right), sum(!human.non.native$right)) out of c(nrow(human.native), nrow(hu
## X-squared = 1064.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.6197722 -0.5402719
## sample estimates:
##     prop 1     prop 2
## 0.03222283 0.61224490
```

With a p-value very much below 0.05, this is a statistically significant difference in the proportion of falsely flagged native vs. non-native essays.

Let's check if this is also true for each individual detector, or if this only happens with some detectors.

```
human.native.grouped <- group_by(human.native, detector) %>% summarise(x1 = (sum(!right)), n1 = n())
human.non.native.grouped <- group_by(human.non.native, detector) %>% summarise(x2 = (sum(!right)), n2 =

getP <- function(x1, n1, x2, n2, name) { # function that just returns the p value and detector name for
  result <- prop.test(x = c(x1, x2), n = c(n1, n2))
  return(c(name, result$p.value))
}
```

```r
ps <- left_join(human.native.grouped, human.non.native.grouped, by = "detector") %>% mutate(p = pmap(li

ps <- ps$p
ps <- as_tibble(do.call(rbind, ps))
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if
## `.name_repair` is omitted as of tibble 2.0.0.
## i Using compatibility `.name_repair`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
print(ps)
```

```
## # A tibble: 7 x 2
##   V1            V2
##   <chr>         <chr>
## 1 Crossplag     2.95010333764005e-06
## 2 GPTZero       1.93808312434779e-13
## 3 HFOpenAI      5.99072468022095e-11
## 4 OriginalityAI 3.25831754188709e-15
## 5 Quil          3.25820982708524e-18
## 6 Sapling       7.51843587120072e-18
## 7 ZeroGPT       2.68833195800247e-12
```

We can therefore see that while the detectors differ in p-values, they are all very much below 0.05, meaning that there is a significant difference in detection accuracy between native and non-native English speakers for all detectors.

To wrap up, we can bookend our notebook with some analysis on the reliability of AI detectors in general.

```r
print(sum(!human.native$right) / nrow(human.native))
```

```
## [1] 0.03222283
```

Interestingly, the incorrect detection rate for native speakers is only roughly 3.2%, making the detector seem somewhat accurate when only considering human written essays for native English speakers. On the other hand, detectors are still likely to incorrectly mark AI-written work as human work roughly 68.8% of the time, meaning that these detectors are basically worse than a coin-flip for detecting whether an AI-generated essay is actually AI or not.

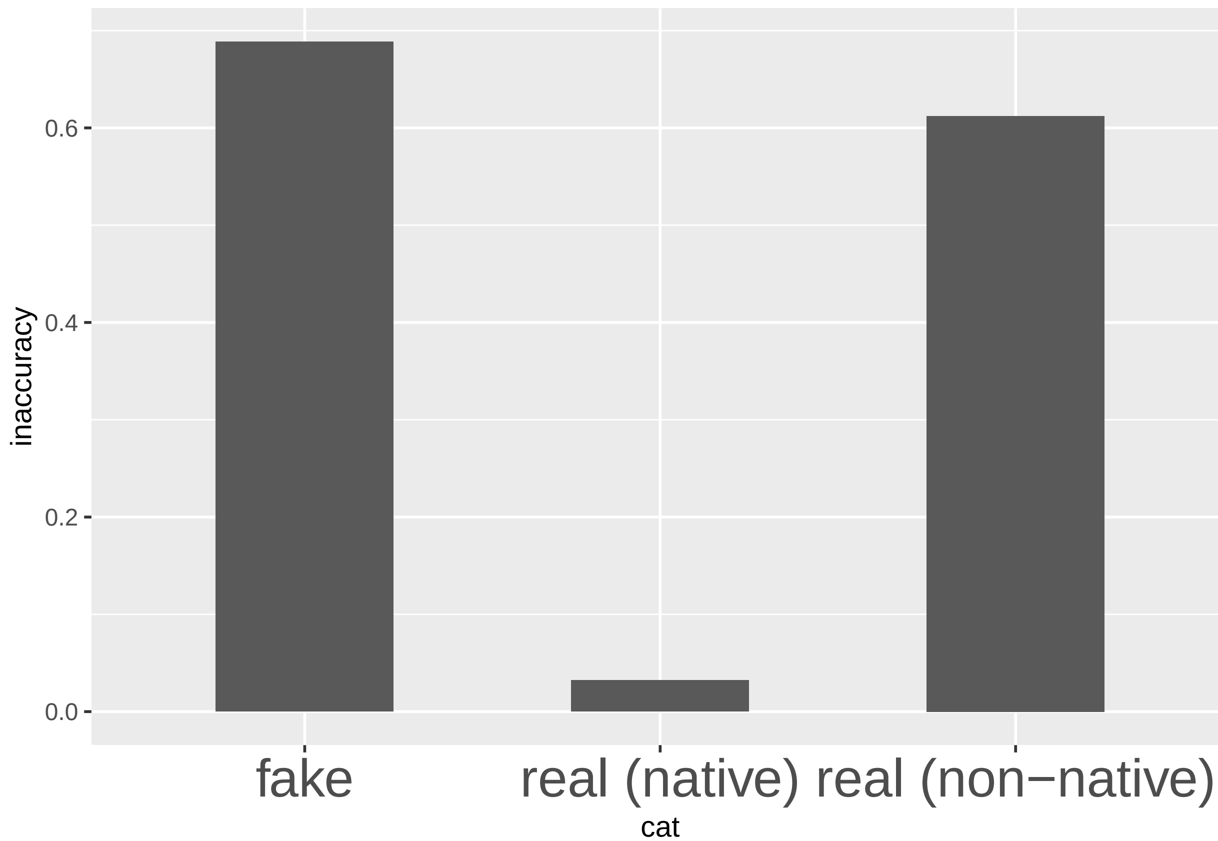
```r
ai.only <- detectors %>% filter(kind == "AI")

print(sum(!ai.only$right) / nrow(ai.only))
```

```
## [1] 0.6884584
```

This can be graphed, too, to compare our three categories of AI, human (native), and human (non-native)

```
final <- detectors %>% group_by(cat) %>% summarise(inaccuracy = (sum(!right) / n()))
ggplot(final, aes(x = cat, y = inaccuracy)) +
  geom_col(width=0.5) +
  theme(axis.text.x  = element_text(size = 20))
```



Conclusions given this data:

In general, AI detectors are currently not reliable ways to detect AI-written work

Across all detectors, human-written essays written by non-native English speakers are much more likely to be incorrectly flagged as AI-generated compared to human-written essays written by native English speakers

AI detectors are only accurate in identifying human-written essays written by native English speakers