Package: detectors (via r-universe)

November 1, 2024

Title Prediction Data from GPT Detectors

Version 0.1.0.9000

Description Researchers carried out a series of experiments passing a number of essays to different GPT detection models. Juxtaposing detector predictions for papers written by native and non-native English writers, the authors argue that GPT detectors disproportionately classify real writing from non-native English writers as AI-generated.

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Depends R (>= 2.10)

LazyData true

URL `https://simonpcouch.github.io/detectors/`, `https://github.com/simonpcouch/detectors`

BugReports `https://github.com/simonpcouch/detectors/issues`

Suggests knitr

Repository https://simonpcouch.r-universe.dev

RemoteUrl https://github.com/simonpcouch/detectors

RemoteRef HEAD

RemoteSha f60679d461c722727536f5b0e3fe1c0a703ae0b0

Contents

`detectors` *Predictions from GPT Detectors*

Description

Data derived from the paper *GPT detectors are biased against non-native English writers.* The study authors carried out a series of experiments passing a number of essays to different GPT detection models. Juxtaposing detector predictions for papers written by native and non-native English writers, the authors argue that GPT detectors disproportionately classify real writing from non-native English writers as AI-generated.

Usage

```
detectors
```

Format

A data frame with 6,185 rows and 9 columns:

kind Whether the essay was written by a `"Human"` or `"AI"`.

.pred_AI The class probability from the GPT detector that the inputted text was written by AI.

.pred_class The uncalibrated class prediction, encoded as `if_else(.pred_AI > .5, "AI", "Human")`

detector The name of the detector used to generate the predictions.

native For essays written by humans, whether the essay was written by a native English writer or not. These categorizations are coarse; values of `"Yes"` may actually be written by people who do not write with English natively. `NA` indicates that the text was not written by a human.

name A label for the experiment that the predictions were generated from.

model For essays that were written by AI, the name of the model that generated the essay.

document_id A unique identifier for the supplied essay. Some essays were supplied to multiple detectors. Note that some essays are AI-revised derivatives of others.

prompt For essays that were written by AI, a descriptor for the form of "prompt engineering" passed to the model.

For more information on these data, see the source paper.

Source

doi:10.1016/j.patter.2023.100779

Examples

```
detectors
```

Index