7. Guerres de Tonner

ChatGPT: Oportunitat i repte per a la docència. I ara què fem?

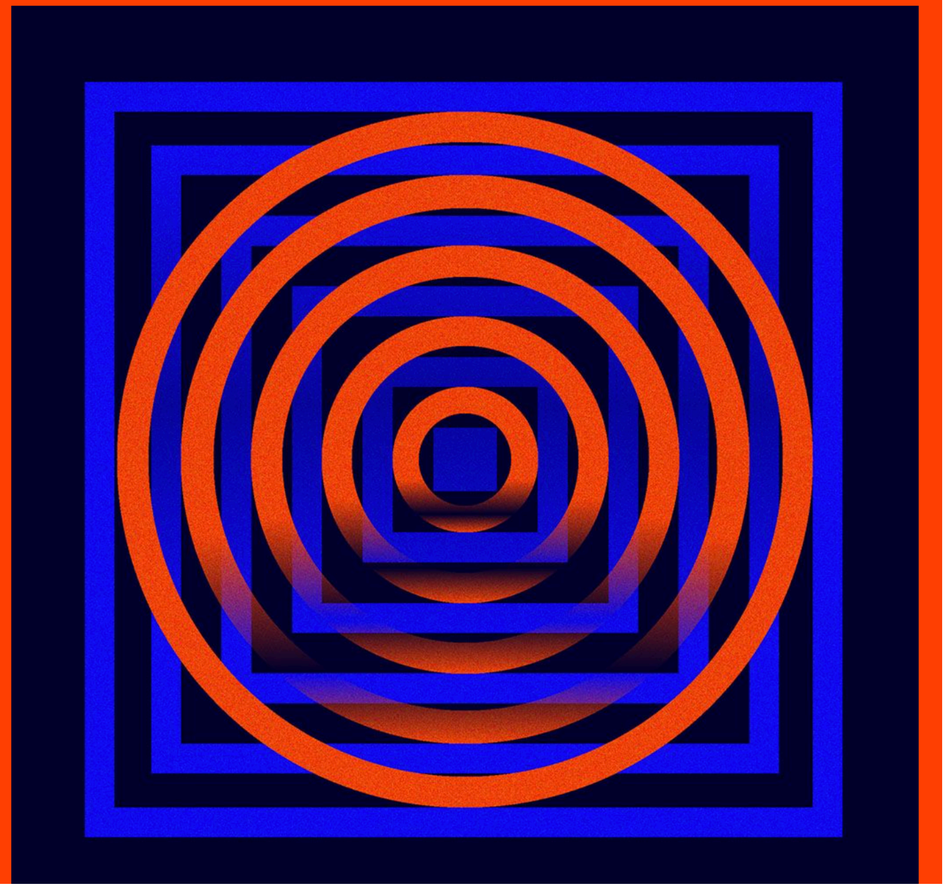
Vídeo 6

Marc Alier

@granludo / ICE - UPC

New AI classifier for indicating AI-written text

We're launching a classifier trained to distinguish between AI-written and human-written text.



January 31, 2023

We've trained a classifier to distinguish between text written by a human and text written by AIs from a variety of providers. While it is impossible to reliably detect all AI-written text, we believe good classifiers can inform mitigations for false claims that AI-generated text was written by a human: for example, running automated misinformation campaigns, using AI tools for academic dishonesty, and positioning an AI chatbot as a human.

Our classifier is not fully reliable. In our evaluations on a "challenge set" of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as "likely AI-written," while incorrectly labeling human-written text as AI-written 9% of the time (false positives). Our classifier's reliability typically improves as the length of the input text increases. Compared to our previously released classifier, this new classifier is significantly more reliable on text from more recent AI systems.

We're making this classifier publicly available to get feedback on whether imperfect tools like this one are useful. Our work on the detection of AI-generated text will continue, and we hope to share improved methods in the future.

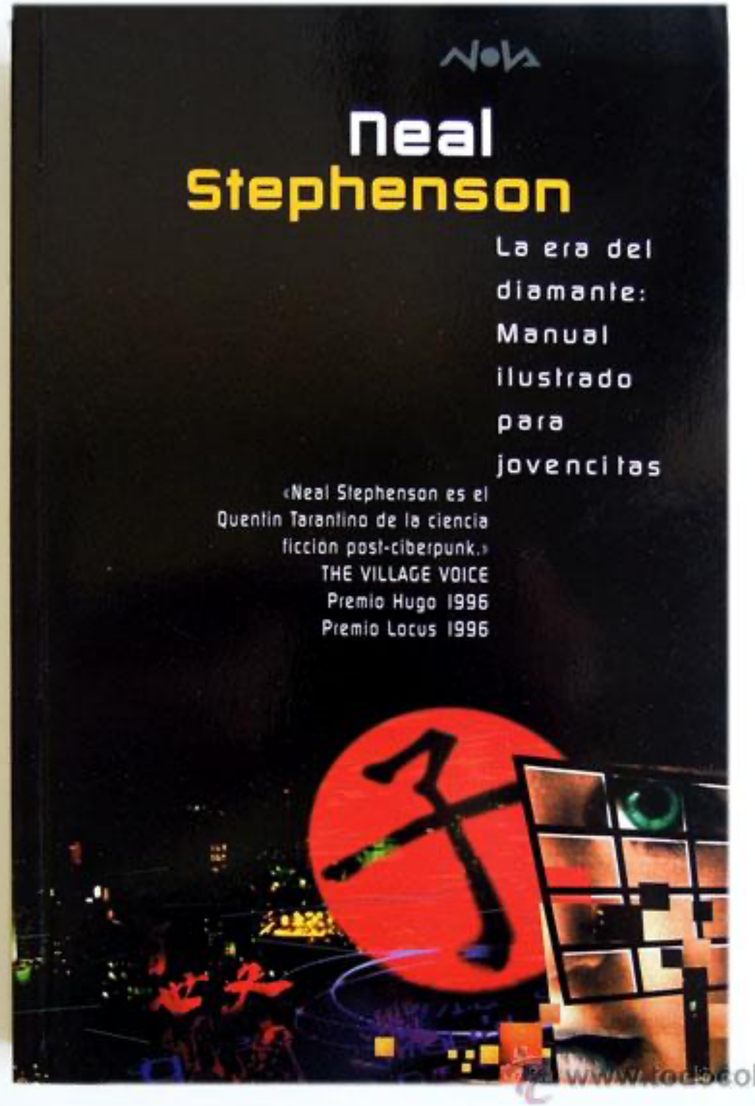
Try our free work-in-progress classifier yourself:

TRY THE CLASSIFIER ↗

Limitations

Our classifier has a number of important limitations. **It should not be used as a primary decision-making tool**, but instead as a complement to other methods of determining the source of a piece of text.

1. The classifier is very unreliable on short texts (below 1,000 characters). Even longer texts are sometimes incorrectly labeled by the classifier.
2. Sometimes human-written text will be incorrectly but confidently labeled as AI-written by our classifier.
3. We recommend using the classifier only for English text. It performs significantly worse in other languages and it is unreliable on code.
4. Text that is very predictable cannot be reliably identified. For example, it is impossible to predict whether a list of the first 1,000 prime numbers was written by AI or humans, because the correct answer is always the same.
5. AI-written text can be edited to evade the classifier. Classifiers like ours can be updated and retrained based on successful attacks, but it is unclear whether detection has an advantage in the long-term.
6. Classifiers based on neural networks are known to be poorly calibrated outside of their training data. For inputs that are very different from text in our training set, the classifier is sometimes extremely confident in a wrong prediction.



arXiv announces new policy on ChatGPT and similar tools

By ame5 January 31, 2023 arXiv updates

The recent release of AI technology that generates new text has raised serious questions among the research community. For one, “Can ChatGPT be named an author of a research paper?”

The resounding answer from arXiv leaders and advisors is, “No.” A computer program cannot, for example, take responsibility for the contents of a paper. Nor can it agree to arXiv’s terms and conditions. Other organizations agree.

To address this issue, arXiv has adopted a new policy for authors regarding the use of generative AI language tools.

The official policy is:

arXiv policy for authors' use of generative AI language tools

January, 31 2023

arXiv recognizes that authors of scientific works use a variety of tools to do the science on which they report, and to prepare the report itself, from simple ones to very sophisticated ones. Community opinion on the appropriateness of such tools may be varied and evolving; AI powered language tools have in particular led to significant debate. We note that tools may generate useful and helpful results, but also errors or misleading results; therefore, knowing which tools were used is relevant to evaluating and interpreting scientific works.

In view of this, we

1. continue to require authors to report in their work any significant use of sophisticated tools, such as instruments and software; we now include in particular text-to-text generative AI among those that should be reported consistent with subject standards for methodology.
2. remind all colleagues that by signing their name as an author of a paper, they each individually take full responsibility for all its contents, irrespective of how the contents were generated. If generative AI language tools generate inappropriate language, plagiarized content, errors, mistakes, incorrect references, or misleading content, and that output is included in scientific works, it is the responsibility of the author(s).
3. generative AI language tools should not be listed as an author; instead authors should refer to (1).


deal with it.