**ESTADO DEL ARTE:**

* **GPTZero:**
  + Xataka (<https://xataka.com/basics/gptzero-que-cuando-usarlo-cuando-funciona-bien-cuando-no>):

En el ámbito educativo, es más fácil de detectar porque los profesores saben cómo escriben sus alumnos.

GPTZero analiza un texto, buscando en el cuerpo del texto fórmulas que se emplean a la hora de expresarse y de escribir contenidos.

Ofrece las posibilidades de detectar textos generados por ChatGPT, GPT-4 o Google Bard; e incluso los creados a medias entre un humano y una IA o solo por humanos.

Devuelve al usuario una ventana donde aparecen los porcentajes de probabilidad de haber sido escrito por la opción escogida de las mencionadas antes.

Único requisito es que el texto contenga entre 250, como mínimo, y 5000 caracteres, como máximo.

Permite escribir/pegar el texto o subir los documentos (formatos .pdf, .doc, .docx y .txt)

No le cuesta identificar si el texto ha sido escrito por una IA o por una persona, lo que le cuesta es identificar qué IA lo ha escrito.

Las limitaciones llegan cuando los textos han sido retocados a mano, sobre todo si son cortos.

* + GTPZero (<https://gptzero.me/news/gpt>):

Las IAs generativas no entienden el concepto del peso humano y emocional de las palabras. Todas las palabras son un cálculo estadístico.

* + GPTZero (<https://gptzero.me/news/how-the-best-ai-detector-provides-interpretable-scores>):

La detección devuelve un porcentaje, el cual hay que saber interpretar y distinguir entre el porcentaje de texto escrito por la IA y la seguridad de que todo el texto esté escrito por la IA.

El detector debería remarcar qué frases cree que están escritas por la IA.

Para no caer en la dicotomía mencionada, se puede mostrar al usuario más resultados de la detección, como los porcentajes de cada categoría (humano, IA o mixto) o el grado de confianza en el resultado dado (baja, media, alta). La tasa de error promedio se marca en un 1% para el cambio de confianza entre alta y media.

* + GPTZero (<https://gptzero.me/news/5-steps-towards-responsible-ai-detection>):

Quizá sea buena idea usar textos en otros idiomas para mejorar la detección, que sea menos sesgada.

* + GPTZero (<https://gptzero.me/news/deep-learning-model-updates>):

El modelo actual es una red neuronal multicapa. En la última actualización, se construyó un pipeline de aprendizaje automático ‘end-to-end’, entrenado con textos masivos de ambos tipos (IA y humano)

* + MediumMMultimedia (<https://www.mediummultimedia.com/apps/como-funciona-el-gptzero/>):

GPTZero (Generative Pre-trained Transformer Zero). Modelo basado en redes neuronales recurrentes con arquitectura Transformer, la cual aprende de manera autónoma a través del proceso de entrenamiento.

Esta arquitectura está compuesta por múltiples capas de atención y autoatención, que permiten al modelo analizar y comprender el contexto.

La autoatención es un aspecto destacado de esta arquitectura, ya que permite al modelo asignar diferentes pesos a las palabras en función de su relevancia para la tarea en cuestión, lo que mejora la coherencia y calidad general del texto generado.

* **Artículos:**
  + <https://arxiv.org/pdf/2301.11305.pdf>:

No necesita entrenar un clasificador separado, ni recopilar un conjunto de datos reales o generados ni usar marcas de agua en los textos (Existen textos generados con un ‘watermark’, o marca de agua, que hace la detección del texto más sencilla)

Se crean copias del texto original modificando frases o partes del este. Luego se comparan los resultados con el texto original (usando una función de probabilidad). Si la probabilidad logarítmica media es alta, significa que el texto original es probable que haya sido generado por el modelo en cuestión.

El enfoque de entrenamiento de redes profundas (como GPTZero) tiene varias deficiencias: la tendencia al sobreajuste o la necesidad de entrenar un nuevo modelo para cada modelo fuente que se publique.

El enfoque ***(Solaiman)*** “zero-shot” a la detección ignora la estructura local de la función de probabilidad aprendida alrededor de un pasaje candidato, que encontramos que contiene información útil sobre el origen de un pasaje.

El texto generado por la IA tiende a ubicarse en zonas de máximos locales en la curva Hessiana de probabilidad logarítmica.

Gráfico

Descripción generada automáticamente con confianza media

El uso de paráfrasis es muy eficiente a la hora de evadir detecciones **(tema para estudiar)**

La detección multilingüe es difícil, se muestra sesgo en contra de los hablantes no nativos **(ver que pasa si se entrena con varias lenguas)**

Texto, Carta

Descripción generada automáticamente



Como hiperparámetros tienen: la fracción de palabras enmascaradas (reescritas por el modelo), la longitud de los textos enmascarados, el modelo para enmascarar y los hiperparámetros de este modelo.

**Trabajos futuros:**

* + - Explorar como se pueden usar los algoritmos de marca de agua junto con algoritmos de detección para mejorar la solidez de la detección.
    - Explorar si usar varios modelos para generar la muestras puede beneficiar el resultado
    - Explorar si la curva de probabilidad aplicada a este proyecto serviría para otros tipos de modelos generativos como audio, imagen…
  + <https://arxiv.org/pdf/2303.11156.pdf>:

Demuestran que el parafraseo es una técnica que impide la correcta detección de muchos modelos detectores.

El parafraseo recursivo utilizado solo degrada ligeramente la calidad del texto, que se mide con estudios en humanos y con métricas como las puntuaciones de perplejidad y la precisión de textos de referencia.

Conectan teóricamente el AUROC del mejor detector encontrado con la distancia de variación total entre las distribuciones de texto humano y de IA. Lo que se puede usar para estudiar el problema de detección fiable para LLMs avanzados

También comentan los problemas de los diversos detectores:

* + - Redes neuronales: Paráfrasis y entrenamiento cada vez que se publica un nuevo LLM
    - Zero-shot: Como indirectamente están basados en la red neuronal del modelo de reescritura, pueden sufrir de ataques adversarios y de envenenamiento **(Goodfellow)**
* Referencias dentro de los artículos a otros artículos:
  + Jawahar, G., Abdul-Mageed, M., and Lakshmanan, L. V. S. Automatic detection of machine generated text: A critical survey. In International Conference on Computational Linguistics, 2020.
  + Solaiman, I., Brundage, M., Clark, J., Askell, A., HerbertVoss, A., Wu, J., Radford, A., and Wang, J. Release strategies and the social impacts of language models, 2019. URL https://arxiv.org/ftp/arxiv/ papers/1908/1908.09203.pdf. **(Enfoque de zero-shot)**
  + Mireshghallah, F., Mattern, J., Gao, S., Shokri, R., and BergKirkpatrick, T. Smaller language models are better blackbox machine-generated text detectors. arXiv preprint arXiv:2305.09859, 2023. **(Mejores funciones de probabilidad cuando no se conoce el modelo a detectar)**
  + Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014. **(Ataques adversarios y envenenamiento)**

**INTRODUCCIÓN**

Los LLM están de moda. La aparición de modelos como Bert **(BERT)** o GPT-2 (OpenAI, 2018) ***(Medium)*** en nuestras vidas ha supuesto una explosión de creación de modelos de lenguaje que tratan de entender el lenguaje humano y de imitarlo, permitiéndonos mantener “conversaciones” con ellos y en los cuales apoyarnos en busca de respuestas o ideas para nuestro día a día.

Esto ha hecho que las personas ganen un “plus” de eficiencia en sus tareas, aunque puede tener inconvenientes si no se contrasta la información que nos brindan, ya que no son perfectos.

De hecho, hoy en día existe una preocupación entre la comunidad de profesorado bastante justificada a mi parecer. Los profesores están preocupados por sus alumnos, dado que estas herramientas están a su alcance y lo más probable es que terminen usándolas para “librarse” de algunas tareas, sobre todo de redacción.

De esta forma, los alumnos no aprenderán como se espera ni lo que se espera, haciendo de estas herramientas algo negativo, además de que la mayoría tampoco contrastarán la información aportada ni reescribirán el texto obtenido.

La idea de este trabajo es intentar captar una tendencia a usar estas herramientas a la hora de generar documentos académicos, creando nuestro propio detector, el cual identificará qué textos, o partes del texto, han sido escritos por alguna de estas Inteligencias Artificiales y cuáles por humanos.

Aplicaremos este detector a textos actuales para sacar conclusiones. Pero antes de comenzar, podemos hacer un trabajo de reflexión y darse cuenta de que, ya que ahora sí que existen estas herramientas, sería de ilusos pensar que no existirá una tendencia alcista de uso en los últimos años.

**ESTADO DEL ARTE**

Actualmente existen algunas herramientas que detectan textos, o partes de estos, generados por alguna IA. El más famoso es GPTZero, un detector creado por Edward Tian, Alex Cui y su equipo de ingenieros y presentado a finales de 2022. ***(Forbes)***

GPTZero es un modelo de redes neuronales recurrentes basado en una arquitectura Transformer. Esta arquitectura está compuesta por múltiples capas de atención y autoatención que permiten al modelo analizar y comprender el contexto de las palabras. ***(MediumMMultimedia)***

La autoatención, o atención propia, es un aspecto que permite al modelo asignar diferentes pesos a las palabras en función de su relevancia, lo que mejora la coherencia y calidad general del texto generado. ***(MediumMMultimedia)***

Está herramienta empezó como detector de ChatGPT, aunque es capaz de detectar textos generados por otros modelos como GPT-4 o Google-Bard. Tras el análisis, GPTZero muestra al usuario un resumen con porcentajes de cada categoría (IA, humano o mezcla) y del grado de confianza en el resultado dado (bajo, medio o alto). ***(GPTZero)***

Otro trabajo investigado es el de DetectGPT **(DetectGPT)**. Este detector obtiene grandes resultados con mucho menos. No necesita entrenar un clasificador por separado, ni recopilar conjuntos de datos, ni la necesidad de marcas de agua en los textos para facilitar la detección.

Utiliza el texto original para crear copias modificadas ligeramente, usando un modelo LLM auxiliar (puede ser el mismo que el detectado) para reescribir alguna frase o parte del texto. Después usan una función de probabilidad logarítmica para comparar estos textos con el original. Si la probabilidad media es alta, significa que el texto original probablemente haya sido generado por el modelo detectado.

Verifican que los resultados de los textos generados por alguna IA suelen ubicarse en las zonas de máximos locales de la función de probabilidad propuesta, mientras que los textos humanos no tienen una posición tan definida.

Gráfico

Descripción generada automáticamente con confianza media

En su artículo, comentan que esta solución es mejor que otras basadas en redes profundas, ya que estas suelen tender al sobreajuste y necesitan entrenar nuevos modelos cuando un nuevo modelo LLM se publica; mientras que su solución solo necesita del texto a detectar y un modelo LLM para las reescrituras. **(Problemas)**

Como problemas remarcan que la paráfrasis obtiene buenos resultados a la hora de evadir deteccionesy que la detección multilingüe es más difícil, ya que se muestra sesgo en contra de los hablantes no nativos. **(Problemas)**

En el trabajo de Sadasivan et al. **(Ataques)** utilizan la técnica del parafraseo recursivo, que obtiene increíbles resultados en contra de los detectores mencionados antes. Incluso pueden con las marcas de agua. Demuestran que crearon un ataque adaptativo al LLM objetivo para aprender su esquema de marca de agua y conseguir que un humano se haga pasar por máquina.

Diagrama

Descripción generada automáticamente

También comentan que, si especializas el detector en algún tipo de fallo el otro crece, ya bien sea Tipo-1 (Falsos positivos, detectar texto humano como IA) o bien Tipo-2 (Falsos negativos, detectar texto generado por IA como humano).

Y concluyen que detectar los textos generados por Inteligencias Artificiales es crítico, pero puede ser mucho peor publicar detectores vulnerables que puedan llegar a incriminar falsamente de plagio a un humano. Si quisiéramos mantener una detección confiable, tendríamos que reducir el rendimiento de los modelos LLM generativos.

**BIBLIOGRAFÍA**

* BERT: <https://aclanthology.org/N19-1423.pdf>
* Medium: <https://medium.com/illumination/the-story-of-chatgpt-and-openai-the-evolution-of-gpt-models-abf201316a9>
* Forbes: <https://www.forbes.com/sites/rashishrivastava/2023/05/09/with-seed-funding-secured-ai-detection-tool-gptzero-launches-new-browser-plugin/>
* GPTZero: <https://gptzero.me/news/how-the-best-ai-detector-provides-interpretable-scores>
* MediumMMultimedia: <https://www.mediummultimedia.com/apps/como-funciona-el-gptzero/>
* DetectGPT: <https://arxiv.org/pdf/2301.11305.pdf>
* Problemas: <https://arxiv.org/pdf/2301.11305.pdf> y <https://arxiv.org/pdf/2303.11156.pdf>
* Ataques: <https://arxiv.org/pdf/2303.11156.pdf>