Detektory wspierające wykrywanie treści wygenerowanych przez AI

Jak działają?

Detektory najczęściej wykorzystują zaawansowane algorytmy uczenia maszynowego. Analizują one cechy charakterystyczne dla tekstów napisanych przez człowieka oraz te, które są typowe dla tekstów wygenerowanych przez sztuczną inteligencję. Często stosowane są różnorodne metody, takie jak analiza stylu pisania, ocena logicznej spójności czy wykrywanie sztucznych wzorców.

Dlaczego są problematyczne?

Główne problemy związane są z oznaczaniem treści jako fałszywie pozytywnych lub fałszywie negatywnych. Fałszywy pozytyw to treść stworzona przez człowieka, ale błędnie oznaczona jako wygenerowana przez maszynę. Z drugiej strony, fałszywy negatyw oznacza treść wygenerowaną przez AI, ale błędnie zidentyfikowana jako stworzona przez człowieka.

Dlaczego jest to trudny problem

Dynamiczny rozwój technologii AI: Technologie generatywne stale ewoluują, co sprawia, że detektory są zazwyczaj opóźnione w stosunku do najnowszych modeli.

Złożoność struktury treści wygenerowanych przez AI: Treści wygenerowane przez maszyny mogą być bardzo zróżnicowane pod względem stylu, tematu i gramatyki. Nie istnieje jednoznaczny wzorzec, który można by wykorzystać do ich identyfikacji.

Adaptacyjność generujących modeli językowych: Generatywne modele językowe, takie jak GPT, są zdolne do uczenia się na podstawie nowych danych i dostosowywania się do nowych kontekstów.

Narzędzia “oszukujące detektory”: Dodatkowym wyzwaniem jest rozwój nowych narzędzi, których celem jest oszukiwanie tych detektorów, i dostosowywanie wygenerowanych treści tak aby wyglądały bardziej “ludzko” (przykładowo https://www.humanizeai.io).

Narzędzia:

Poniższa lista narzędzi stanowi pewien wybór, i nie pokrywa wszystkich dostępnych na rynku rozwiązań. Sytuacja w świecie detektorów jest dosyć dynamiczna, mamy wiele projektów na różnym etapie rozwoju, nie wszystkie z nich przetrwają lub będą użyteczne wraz z upływem czasu.

Detektory w języku angielskim:

1. **GLTR** (Giant Language Model Test Room): Narzędzie open-source do wykrywania tekstu wygenerowanego przez zaawansowane modele językowe, analizujące prawdopodobieństwo kolejnych słów. Link: http://gltr.io/
2. **Detector** od OpenAI: Narzędzie OpenAI służące do identyfikacji treści wygenerowanych przez różne modele językowe. Jego skuteczność może się zmieniać z czasem. Link: https://platform.openai.com/ai-text-classifier
3. **ZeroGPT**: Model specjalizujący się w wykrywaniu tekstu wygenerowanie przez ChatGPT i GPT-4, dostępny w różnych opcjach cenowych, z łatwym dostępem poprzez konto użytkownika. Jest dostępna wtyczka do przeglądarki Chrome. Link: https://www.zerogpt.com/
4. **GPTZero** (mimo podobieństwa w nazwie do poprzedniego, to inne rozwiązanie): narzędzie do wykrywania treści AI, w tym ChatGPT i GPT-4, oferujące różne rozwiązania dla nauczycieli, studentów, pisarzy. Link: https://gptzero.me/
5. **Radar Tester** (Robust AI-Text Detection via Adversarial Learning): To narzędzie służy do wykrywania tekstu wygenerowanego przez sztuczną inteligencję, wykorzystując techniki uczenia przeciwnego (adversarial learning). Ma na celu zwiększenie skuteczności identyfikacji AI-generowanych treści poprzez adaptację do ewoluujących metod generacji tekstu.: Link: https://radar-app.vizhub.ai/
6. **Free AI Content Detector** - Bezpłatny detektor treści AI umożliwia szybkie wykrywanie zawartości wygenerowanej przez AI, takich jak ChatGPT 3.5, GPT-4 i Google Bard. Link: https://www.scribbr.com/ai-detector/
7. **Copyleaks AI Content Detector**: Detektor treści AI dla popularnych modeli. Link: https://copyleaks.com/ai-content-detector
8. **Writer Ai detector**: Link: https://writer.com/ai-content-detector/
9. **AISEO**: Link: https://aiseo.ai/AI-tools/AI-text-detector.html

Detektory w języku polskim:

1. **Monica**: Link: https://monica.im/tools/ai-content-detector?locale=pl
2. **Smodin**: Link: https://smodin.io/pl/detektor-tresci-ai
3. **textflip**: Link: https://textflip.ai/pl/ai-writer-detector/

Nowe narzędzia będą dodawane do listy. Warto zapoznać się też z innymi listami (Giles, L.).

Podsumowanie

W naszej opinii, ze względu na wskazane wyżej problemy, wynik działania detektora może być pewną podpowiedzią dla prowadzącego, na tę chwilę jednak nie może stanowić jedynego mechanizmu podejmowania decyzji.

Na co zwracać uwagę przy weryfikacji:

1. “Halucynacje AI” - zjawisko, w którym systemy sztucznej inteligencji generują treści, które wydają się być autentycznymi, ale nie mają odpowiedniego odniesienia w rzeczywistości.
2. Wynik działania detektorów - wraz z krytyczną oceną wyjaśnień, które detektor zwraca (oznaczone konkretne frazy czy fragmenty obrazów).
3. W przypadku wątpliwości, bezpośrednią rozmowę ze studentem.

Literatura

- Giles, L., AI Detector Tools: The Best and The Worst [WWW Document]. URL https://www.linkedin.com/pulse/ai-detector-tools-best-worst-lester-giles (accessed 2.14.24).