EFCSN

Jak zidentyfikować treści zmodyfikowane cyfrowo i wygenerowane przez sztuczną inteligencję

Łatwo dostępne i proste w obsłudze modele sztucznej inteligencji pomagają w nauce i tworzeniu nowych treści. **Mogą one jednak zwiększyć ryzyko rozprzestrzenienia się dezinformacji i błędnych informacji** w społeczeństwach otwartych i debacie demokratycznej. Należy dbać o to, żeby wspólnie używana przez nas "przestrzeń internetowa" nie została zapełniona dezinformacjami wygenerowanymi przez sztuczną inteligencję, albo błędnymi informacjami wcześniej zmodyfikowanymi cyfrowo.

Możemy tego uniknąć, stosując nowe technologie, np. programy poświadczające pochodzenie treści cyfrowych, i oprogramowania służące do wykrywania treści wygenerowanych przez AI. **Jednak rozwiązaniom technicznym jeszcze daleko do ideału i to niezależni weryfikatorzy zapewniają i udostępniają społeczeństwu sprawdzone informacje.**

Oto co robią profesjonalni, niezależni weryfikatorzy, żeby zidentyfikować dezinformacje i błędne informacje oraz czego można się od nich nauczyć.

Coraz więcej treści generowanych przez SI

Dezinformacje i błędne informacje wygenerowane przez sztuczną inteligencję to bardzo mały odsetek danych sprawdzanych przez profesjonalnych, niezależnych weryfikatorów, którzy głównie zajmują się weryfikacją i wyszukiwaniem treści zmodyfikowanych cyfrowo.

Jednak w międzynarodowej ankiecie przeprowadzonej przez EFCSN (Europejska Sieć Standardów Fact-Checkingowych) większość weryfikatorów zgodziła się, co do tego, że treści generowanych i modyfikowanych przez sztuczną inteligencję będzie coraz więcej.

WAŻNE: *Zmodyfikowany cyfrowo* odnosi się do treści, które zostały znacząco przerobione w celu manipulacji lub, którym nadano inne znaczenie, a także do tych treści, które zostały zedytowane przez narzędzia SI. Nie obejmuje to treści, zedytowanych w celu poprawienia jakości, czy wyrazistości (obrazu, dźwięku, etc).

Wygenerowany przez sztuczną inteligencję odnosi się do treści, które zostały stworzone przez program SI.

Informacje zostały zebrane przez **AI@EUElections,** projekt finansowany przez firmę Meta i prowadzony przez **Europejską Sieć Standardów Fact-Checkingu** i jej członków oraz zewnętrznych specjalistów.

Technologia szybko się rozwija, ale nie możemy na niej polegać.

Specjaliści do spraw sztucznej inteligencji i profesjonalni weryfikatorzy zgadzają się co do tego, że **programy poświadczające pochodzenie treści cyfrowych nie wystarczają, by zidentyfikować treści zmodyfikowane cyfrowo i te wygenerowane przez sztuczną inteligencję.**

Przed użyciem takiego programu, zalecane jest, zapoznanie się z działaniem różnych narzędzi **generujących treści** za pomocą SI. Dzięki zrozumieniu, w jaki sposób trenuje się modele SI oraz podstawowym znajomościom statystyki, weryfikatorzy mogą rozpoznać mocne i słabe strony, a także przyszłość takich narzędzi.

Poświadczanie treści
Inicjatywy takie jak C2PA mają na celu sprawdzenie źródła i historii treści udostępnianych w internecie. Niestety różnego rodzaju oznakowania, np. znaki wodne też można podrobić.

Jak dezinformacja spowodowana przez sztuczną inteligencję wpływa na ludzi

"Za każdym razem gdy ulegasz instynktowi, to jakby to powiedzieć, załamuje to twój obraz rzeczywistości."
– Christine Dugoin*

Operacje wpływu są często zaplanowane tak, żeby wykorzystać błędy poznawcze.

Zrozumienie błędów popełnianych samemu i takich, jakie może popełnić ktoś z docelowej grupy odbiorców, pomaga w identyfikacji dezinformacji

Prowokatorzy, przestępcy i spiskowcy mogą polegać na SI i generować oraz rozprzestrzeniać dezinformacje. Jaki może być skutek ich działań?

- ➔ mogą rozszerzyć swoje wpływy na inne kraje lub społeczności
- ➔ zbyt duża ilość podobnych fałszywych lub błędnych informacji może przytłoczyć weryfikatorów
- ➔ taka dezinformacja może zostać udostępniona przez kilka podszywających się kont i zyskać wiarygodność wśród odbiorców

** Christine Dugoin - badaczka na Uniwersytecie Sorbońskim, specjalizująca się w badaniach dotyczących wpływu informacji na człowieka*

Weryfikacja wymaga wielopoziomowego i szczegółowego podejścia

Skoro programy poświadczające pochodzenie treści cyfrowych nie działają, to co działa? Profesjonalni weryfikatorzy mają bardzo dobrze rozwinięte umiejętności śledcze. Oto kilka wskazówek od ekspertów:

> *"Programy poświadczające pochodzenie treści cyfrowych nie są w 100% skuteczne - I pewnie nigdy nie będą."*
> - Henk van Ess**

CZY ŹRÓDŁO INFORMACJI JEST WIARYGODNE: Czy można ustalić tożsamość twórcy? Czego dotyczą zwykle udostępniane przez niego treści? Co można powiedzieć o użytkownikach, którzy wchodzą w interakcję (np. komentują) z autorem? Jaki wpływ mogą mieć ukazane treści na odbiorcach?

USTAL WIARYGODNOŚĆ: Niezależnie zweryfikuj podane informacje w rzetelnych źródłach, np. Zobacz czy specjaliści w danej dziedzinie wypowiedzieli się na ten temat . Does what is depicted make sense based on your knowledge?

UŻYJ TECHNIK ŚLEDCZYCH: Poza klasycznymi sposobami szukania informacji i wyszukiwaniem badań naukowych, skorzystaj z technik śledczych takich jak m.in.: data scraping, szukanie geolokalizacji, rozpoznawanie biometryczne, analiza danych itp.

NAUKA I DOSKONALENIE: Generowane przez sztuczną inteligencję dezinformację stają się coraz trudniejsze do rozpoznania. Dlatego pamiętaj o ciągłym doskonaleniu swoich umiejętności.

PODZIEL SIĘ SWOJĄ PRACĄ Z INNYMI

Zidentyfikowanie dezinformacji to nie wszystko - zaleca się stworzenie przejrzystej analizy wraz z linkami i odnośnikami do źródeł. Dzięki niej czytelnicy z łatwością będą mogli śledzić dochodzenie i ocenić jego wiarygodność.

*** Henk van Ess jest specjalistą w OSINT (biały wywiad) i technikach weryfikacji informacji.*

Skrócona instrukcja obsługi: Zalecenia, ostrzeżenia i wskazówki

Poniższe wskazówki mogą okazać się pomocne w wyszukiwaniu zmodyfikowanych cyfrowo i wygenerowane przez sztuczną inteligencję treści. Wraz z innymi poradami zawartymi w tym krótkim poradniku (zwrócenie uwagi na kontekst, techniki śledcze i programy poświadczające treści cyfrowe), pomogą Ci odnaleźć prawdę.

tekst

- Często (lecz nie zawsze) **poprawnie napisany** (właściwa składnia i gramatyka)
- **Zbyt formalny lub zbyt skomplikowany**, zwłaszcza w kontekście mediów społecznościowych
- **Nadmierna liczba przymiotników i przysłówków**.
- Bez emocji, humoru, ironii, czy wyrażeń idiomatycznych
- **Nieszczegółowy** (brak imion, dat, miejscowości itd.)
- Najważniejsze: czy fakty przytoczone w tekście są prawdziwe?

nagranie dźwiękowe

- **Porównaj podejrzane nagranie dźwiękowe z autentyczną próbką,** użyj narzędzi, które wyszukują różnice w mowie, sposobie oddechu, i intonacji...
- Używając tych narzędzi unikaj nagrań o słabej jakości lub z szumami i zakłóceniami w tle
- Zwróć uwagę na **nienaturalne lub nadmiernie monotonne sposoby mówienia**, brak przerw lub naturalnych odetchnienia.

nagranie wideo

- Nie używaj programów służących do detekcji wygenerowanych przez SI obrazów na klatkach filmowych.
- Przyjrzyj się **wyrazom i ruchom twarzy**, np. Czy osoby na nagraniu mrugają i czy słowa przez nich wypowiadane pasują do ruchu ich ust?
- Czy przejścia do kolejnych klatek są łagodne, czy można zauważyć **ostre cięcia?**

zdjęcia

- Poszukaj **nienaturalnych szczegółów**: idealna skóra, zamazane tło, nienaturalne krajobraz lub światło i różnego rodzaju osobliwości np. dodatkowy palec.
- **Poszukaj znaku wodnego** z generatora obrazów
- Zwróć uwagę na szczegóły: czy to logiczne? Czy to poprawne?
- Podczas korzystania z detektorów, **wybierz wersję obrazu o wysokiej rozdzielczości lub pierwotną wersję obrazu, a nie taką, która została udostępniona wielokrotnie**.