



AI en Media

Beschrijving

Online desinformatie (fake news) die politiek geweld kan aanwakken en verkiezingen kan verstoren, is een groot probleem van onze tijd. Technologen en beleidsmakers bespreken dit uitgebreid, maar hebben vooralsnog geen idee hoe het echt aangepakt moet worden. Dit heeft te maken met de complexe wisselwerking tussen

- de omvang van sociale media-platforms,
- de opkomst van geavanceerde tools voor het bewerken en synthetiseren van content (AI) en
- de aard van menselijke communicatie.

Degenen die op zoek zijn naar mogelijke oplossingen, hebben de plank misgeslagen. Niettemin zijn tools die ook worden aangestuurd door AI een kanshebber gebleken om te proberen bewijs van bedrog te detecteren in een breed scala aan content.

AI-detectie

[TrueMedia.org](https://www.truemedia.org/), een non-profitorganisatie die zich toelegt op het detecteren van deepfake- of kunstmatige video's, pocht dat haar detectietool 90 procent nauwkeurig is.

Een recent productaanbod van start-up *Sensity AI* dat valse video's, afbeeldingen, audio en identiteiten detecteert, beweert 98 procent nauwkeurig te zijn.

Intel, de halfgeleidergigant, is ook in deze arena gegaan met zijn *FakeCatcher*, een AI-algoritme dat is ontworpen om te meten wat Intel â??subtiële bloedstroomâ?? noemt bij mensen die in video's worden afgebeeld.

De website van het bedrijf beweert: â??Wanneer ons hart bloed pompt, veranderen onze aderen van kleur. Deze bloedstroomsignalen worden verzameld van over het hele gezicht en algoritmen vertalen deze signalen naar spatiotemporele kaarten. Vervolgens kunnen we met behulp van deep learning direct detecteren of een video echt of nep isâ??. Intel zegt dat het 96 procent nauwkeurig is.



Peter en de Wolf¹

Hoewel dit opwindende beweringen zijn, moeten gebruikers hier voorzichtig zijn. Het is belangrijk om te begrijpen dat de voorgestelde oplossingen nieuw zijn en nog nauwelijks zijn getest door derden in de complexe werkelijke online wereld. Ontwikkelaars roemen doorgaans alleen de sterke punten van hun vinding en niet de fundamentele beperkingen. Door geen rekening te houden met de beperkingen van de huidige technologie voor het detecteren van desinformatie, kunnen gebruikers de plank volledig mislaan en net als in het sprookje van *Peter en de Wolf* **te vaak** *“wolf roepen”* iets dat misschien net zo gevaarlijk is als de desinformatie zelf.

Desinformatie neemt verschillende vormen aan

Een deel van de desinformatie is tekst-gebaseerd. Een voorbeeld hiervan is traditioneel *“nepnieuws”*, of valse informatie die wordt gerapporteerd alsof het waar is.

Misleidende content kan ook de vorm aannemen van stilstaande beelden. Een veelvoorkomend probleem met afbeeldingen is dat ze worden bewerkt of gesynthetiseerd om de geschiedenis te herzien op een manier die een reactie van de waarnemer zal uitlokken.

Sommige misinformatie kan in de vorm van audio zijn. Recente oplichtingspraktijken zijn gericht op ouderen. Het zijn telefoongesprekken waarin AI zich voordoeft als familielid.

Multimedia, zoals een *deepfake* videopost op sociale media die zowel beeldmateriaal als audio combineert om een â??echt persoon te simuleren, zijn natuurlijk ook nep.

Simpele patronen

Kunstmatige intelligentietools kunnen al deze signalen afzonderlijk of samen onderzoeken, maar hoe meer signalen erbij betrokken zijn, hoe moeilijker de verwerking wordt. De meeste AI-detectoren vertrouwen op machinaal leren (het gebruik van training met voorbeeldgegevens om automatisch patronen te leren) om voorbeelden van verschillende categorieën desinformatie te identificeren. De patronen die de detectoren leren zijn doorgaans simplistisch:

- inconsistenties in de statistieken van de pixels in een afbeelding,
- het herhaaldelijke gebruik van bepaalde woorden of zinnen in tekst, of
- niet-overeenkomende afbeeldingen en bijschriften.

AI tegen AI, wie zal er winnen?

1 **Peter en de wolf** is een muzikaal sprookje van *Sergej Prokofjev* voor spreekstem en orkest. Het stuk ging in 1936 in Moskou in première en is wereldwijd zeer succesvol geworden.

Datum aangemaakt

2024/10/22