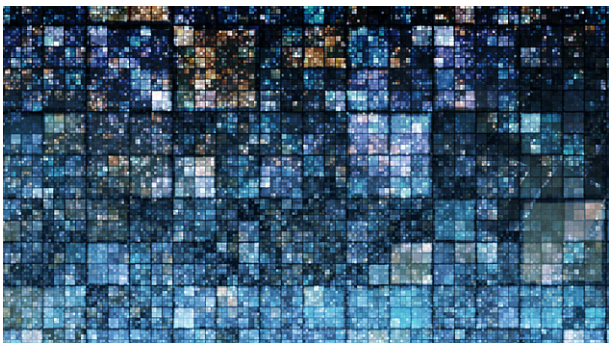
A moderálás egyelőre nem a mesterséges intelligencia erőssége

Egy új kutatás szerint a legjobb gépi tanuló modellek is csak küszködnek a gyűlöletbeszéd meghatározásával, sokszor éppen úgy, hogy egymás után dobálják a fals pozitívokat.



A gyűlöletbeszéd, vagyis a társadalmi csoportok megalázását, megfélemlítését és ellenük való erőszakos fellépés kiváltását célzó megnyilvánulás az online térben is meghatározó probléma, ennek kezelésére pedig a korábbi várakozásokkal ellentétben a legfejlettebb gépi tanuló modellek sem hoztak megfelelően működő, automatizált megoldást. Ennek oka elsősorban az, hogy a természetes nyelveket feldolgozó technológiák szempontjából is különösen összetett dologról van szó, amelynek számtalan megnyilvánulási formáját nehéz megkülönböztetni a józan ész szerint ártalmatlannak tekinthető közlésektől.

Az algoritmikus moderáció természetesen így sem teljesen sikertelen: a Facebook november végi közlése szerint például a megelőző negyedévben több mint 22 millió esetben lépett fel ilyen tartalmak ellen, és a megjelenések 95 százalékát proaktív módon azonosította, vagyis az MI segítségével még azelőtt megjelölte vagy el is távolította azokat, hogy felhasználói bejelentés érkezett volna velük kapcsolatban. Ez persze nem jelenti, hogy egy nagy csomó üzenet nem jutott át a szűrőn és nem terjedt tovább a kívánatosnál sokkal szélesebb körben, mint ahogy az is igaz, hogy a rendszer időnként olyasmit is megfog, ami teljesen nyilvánvalóan kívül esik a gyűlöletbeszéd vagy a felhasználási feltételeket másképpen sértő tartalmak körén.

A szóban forgó modellek eredményességét hagyományosan az olyan metrikák alapján értékelik, mint mondjuk a megfelelő mintákon értelmezett pontosság, ami azonban nem segít a gyenge pontok azonosításában, sőt a felhasznált adatkészletek hiányosságai miatt sokszor a modellek minőségének túlértékeléséhez vezethet. A megfelelő benchmark kidolgozása ezért alapvető lenne ezeknek az alkalmazásoknak a továbbfejlesztéséhez. Erre kínálna most megoldást a HateCheck nevű, első körben 29 funkcionális tesztből álló készlet, amit az Oxfordi Egyetem, a Sheffieldi Egyetem, az Utrechti Egyetem és a londoni Alan Turing Intézet kutatói közösen állítottak fel a rendszerek teljesítményének összehasonlítására.

Az egyelőre csak angol nyelvű benchmark csomagba összesen 16, az online gyűlöletbeszédet kutató amerikai, brit és német NGO eredményeit is beépítették. A HateCheck a fenti publikáció szerint több csúcskategóriás modell kritikus gyengeségeit is feltárta, ami a kutatók szerint máris igazolta a hasznosságát. A tesztkörnyezet állítólag alapból is nehéz feladat elé állítja azokat az algoritmusokat, amelyek leegyszerűsített szabályokat alkalmaznak: a 29 tesztből 18 valóban a derogatív és fenyegető kifejezések, illetve a gyűlöletbeszédet kísérő trágárság világos megnyilvánulásairól szól, a másik 11 azonban az a kontrasztív elemzést, esetleg a gyűlöletbeszédre is jellemző lingvisztikai sajátosságokat próbálja lefedni.

Az elfogultság egyelőre a rendszerekbe van kódolva

A kutatók úgy találták, hogy mindegyik most vizsgált modell a kelleténél érzékenyebb bizonyos kifejezésekre, miközben gyakran osztályozza rosszul a gyűlöletbeszédnek nehezen minősíthető szembeállításokat (tagadás, ellenbeszéd). Egyes modellek-

nek azzal is nagyon komolyan meggyűlik a baja, ha a gyűlöletbeszédet denunciáló bejegyzések idézetet vagy hivatkozást tartalmaznak, vagy ha a gyűlöletbeszéd olyan csoport ellen irányul, ami máskülönben nem számít az ilyen támadások megszokott célpontjának. Az egyes csoportok között egyébként is jól mérhető különbségek vannak: a nők vagy a mozgássérültek ellen szóló gyűlöletbeszédet például jelentősen kisebb eséllyel szűrik ki a vizsgált modellek, mint ha a bevándorlókról vagy a feketékről lenne szó.

A kutatók szerint világos, hogy még a legfejlettebb gépi tanuló technológiák is többé-kevésbé az egyszerű, kulcsszavakra épülő döntéshozatalt alkalmazzák a releváns lingvisztikai jelenségek azonosítása helyett. Ezen felül képtelenek jól elkülöníteni azokat a nyelvi jeleket, amelyek a gyűlöletbeszédet képviselő mondatokból éppen hogy ellentétes értelmű közléseket faragnak. Ahogy a VentureBeat beszámolójából is kiderül, itt természetesen hasznos lenne az egyes modellek továbbtanítása olyan adatkészleteken, amelyeket a felfedezett gyengeségek alapján állítanak össze, bár ez még nem oldana meg egy másik fontos problémát, mégpedig a kifogásolt tartalom terjedésének kontrollját.

Ahogy a szerzők az NBC egy nemrég közölt oknyomozó anyagára hivatkozva megállapítják, a Facebook Instagram szolgáltatásában a fekete felhasználók fiókjait arányaiban 50 százalékkal gyakrabban függesztik fel az automata moderációs rendszerek. Ezt azonban nem csak úgy lehet értelmezni, hogy velük szigorúbb, hanem úgy is, hogy a fehérekkel szemben megengedőbb; a lényeg azonban mindenképpen az, hogy a gyűlöletbeszédnek minősített tartalmak terjedése az egyes környezetekben sokkal simább a többihez képest. Az automatizált moderálásról szólva ezek olyan kritikus hiányosságok, amelyek éppen azokat a rendellenességeket szilárdítják meg vagy teremtik újra, amelyek ellen a technológiát eredetileg alkalmazni akarták.

Forrás: https://bitport.hu/a-moderacio-egyelore-nem-a-mesterseges-intelligencia-erossege

Válogatta: Fonyó Istvánné