Finoman módosított szövegekkel átverhető a mesterséges intelligencia

A szöveg alapú MI-modellek sérülékenyek lehetnek az okosan kiválasztott parafrázisokkal szemben, amelyek az emberek szemében nem okoznak jelentésváltozást, a gépi algoritmust viszont simán megvezetik.

A természetes nyelvek feldolgozása (NLP) a mesterséges intelligencia más területeihez hasonlóan jelentős fejlődésen ment keresztül az elmúlt években, így az sem csoda, hogy a vállalatok és különböző szervezetek egyre nagyobb arányban használnak MI algoritmusokat az olyan, szöveg alapú feladatok támogatására, mint amilyen mondjuk a levélszemét kiszűrése, a közösségi média és a vásárlói értékelések véleményelemzése, vagy akár az álhírek filterezése a különböző csatornákon.

A szóban forgó algoritmusok egyre megbízhatóbb munkát végeznek, így az automatizálás ebben a tekintetben is kifizetődőnek tűnik. Egy friss kutatás azonban felhívja rá a figyelmet, hogy sebezhetőségekkel ezen a területen is számolni kell: az IBM, az Amazon és a University of Texas közös vizsgálata szerint a rosszindulatú szereplőknek megvannak az eszközeik a szöveges tartalmat osztályozó rendszerek támadására, amelyekkel hatékonyan befolyásolhatják azok működését.

Addig fogalmazgatják, amíg egyszer át nem csúszik

Az eredményekről a Stanford április elsején rendezett SysML AI konferenciáján számoltak be, parafrazeáló támadásnak (paraphrasing attack) nevezve azt a módszert, amellyel a bevitelre váró szövegeket úgy módosítják, hogy azok jelentése ne változzon érdemben, a gépi intelligencia viszont homlokegyenest másképp osztályozza azokat. Vagyis egy spam üzenetet például úgy juttassanak át csont nélkül a szűrőn, hogy annak tartalma ugyanaz maradjon a címzett olvasatában.

A kép- vagy hangfelismerő algoritmusokat már eddig is hasonló módon támadták, olyan változtatásokat eszközölve az eredeti anyagokon, amelyek az MI-t átverték, de a tartalomfogyasztók szemében nem számítottak zavarónak. Ahogy azonban a képpontok színének fokozatos átkeverésével ki lehet tapasztalni, hogy mi csapja be a szűrőt, úgy a kutatók a sebezhetőségeket is modellezni tudják. A diszkrét szöveges állományok esetében a támadóknak is nehezebb a dolga, hiszen nem próbálkozhatnak olyasmivel, hogy 10 százalékkal többször írják bele a „kutya” szót a szövegbe, aztán megnézik, hogy mi lesz a dolog vége. Ezzel párhuzamosan viszont a védekezés is bonyolultabb, pontosan azért, mert nehéz tipizálni és modelleket állítani a sérülékenységekre.

A mostani kutatás mögötti ötlet éppen erre épül: ha sikerülne szintén a mesterséges intelligencia segítségével feltárni a gyenge pontokat, akkor célzottan fel is lehetne lépni a rosszindulatú kísérletekkel szemben. Ez annál is fontosabb, mivel a szöveges manipulációk hagyományosan egy-két megfelelő szó cseréjére épülnek, ez azonban sok esetben értelemzavaró, és mesterséges hatást kelt az emberi befogadóknál. A parafrazeáló támadás viszont egész mondatokat cserél ki (gyakran sokkal hosszabb mondatokra), így az értelmezés nem sérül, csak a szűrő képtelen kezelni az új megfogalmazást.

Parafrazeáló sárkány ellen parafrazeáló sárkányfű

A kutatók által fejlesztett algoritmus is parafrazeál: egy-egy kiválasztott mondat mellé szemantikusan hasonló szekvenciákat generál, és megnézi, hogy a vizsgált technológiák ugyanúgy értékelik-e az új mondatokat, mint az eredetit. A rendszer azokat az optimális változtatásokat keresi, amelyek eltérítik az NLP modellek működését: szélsőségesen kitágítja a szinonimák és parafrázisok keresési tartományát, kiválasztja a leghatékonyabb változatot, elméletileg igazolja a választást, és az automatizálás révén alaposan fel is gyorsítja ezt az időigényes folyamatot.

A dolog érdekessége, hogy az emberi felhasználók gyakorlatilag képtelenek lennének kiszűrni a parafrazeáló támadásokat, pont azért, mert nincs jelentésbeli különbség, és az így előállított szövegek sem hatnak idegenszerűnek – ezt a szakemberek kísérletekkel is igazolták. A gépekkel persze más a helyzet. Az emberek kevésbé érzékenyek a koherenciára, mivel naponta ezerszer találkoznak tökéletlen inputokkal, vagyis alapból nem kezdenek egy háttérben dolgozó algoritmusra gyanakodni.

A gépi intelligencia viszont nem így működik, és a tanulmány szerint lassan ideje lenne komolyan foglalkozni a problémával, ahogy a szöveges állományok osztályozásában a szoftverek is egyre nagyobb szerepet kapnak. A vállalati IT-fejlesztések elsősorban az automatizációra és a skálázhatóságra fókuszálnak, közben a biztonságra nem allokálnak elegendő forrást – különös tekintettel az ilyen, egyelőre nem kézzel fogható kockázatokra. A mostani kutatás viszont azt támasztja alá, hogy a parafrazeáló támadások MI alapú visszamodellezése hatékony eszközt jent a védekezésben, pontosabbá és általánosabbá téve a vonatkozó biztonsági készségeket.

Forrás: https://bitport.hu/finoman-modositott-szovegekkel-atverheto-a-mesterseges-intelligencia

Válogatta: Fonyó Istvánné