



## Kako delujejo orodja za zaznavanje besedil ustvarjenih z umetno inteligenco

Avtorja: Jani Dugonik, Domen Verber, Univerza v Mariboru – FERi

Generativna umetna inteligenca je v zadnjih letih dobesedno spremenila način, kako nastajajo besedila. Orodja, kot je ChatGPT, danes z nekaj kliki napišejo eseje, seminarske naloge, oglase ali celo članke, kot je ta. Takšna orodja znajo posnemati različne sloge pisanja, uporabljajo ogromne količine podatkov in vedno bolj zvenijo "človeško". Hitrost, s katero se ta tehnologija razvija, pa prinaša tudi dilemo: kdo je resnični avtor besedila, ki ga beremo? V dobi, ko je mogoče v nekaj sekundah ustvariti prepričljiv esej ali prispevek, se povečuje tudi potreba po preverjanju avtentičnosti. Zato so se razvila orodja, imenovana detektorji, ki skušajo zaznati, ali je besedilo napisal človek ali umetna inteligenca. Večina takšnih detektorjev temelji na merjenju perpleksnosti. Veliki jezikovni modeli pišejo gladko, urejeno in po nekakšni statistični logiki, medtem ko je človeško pisanje pogosto bolj nepopolno, z napakami, stilskimi posebnostmi, celo protislovji.

Detektorji zato iščejo vzorce, značilne za umetno inteligenco, kot so ponavljanje istih besed, bolj pravilne stavčne strukture ter odsotnost osebnih vtisov ali pristnih čustev. Obstajajo štirje pristopi za odkrivanje vsebin ustvarjenih z umetno inteligenco (slika). **Watermark** v besedilo vdela posebne oznake, ki jih detektor prepozna. **Zero-shot** uporablja statistične analize izhodov modelov za odkrivanje brez dodatnega učenja. **Fine-tuning LMs** temelji na posebej prilagojenih jezikovnih modelih, ki so naučeni razlikovati med človeškim in besedilom ustvarjenim z umetno inteligenco. **LLMs as Detector** pa preprosto sprašuje velike jezikovne modele, ali je besedilo ustvaril človek ali umetna inteligenca.



Slika 1: Tehnike zaznavanja.

Med najbolj znanimi detektorji so GPTZero, ZeroGPT in Copyleaks. Njihova uporaba je preprosta: besedilo prilepimo v vmesnik, orodje ga analizira in poda verjetnost, da ga je napisal stroj. Takšna orodja uporabljajo predvsem šole, univerze, uredništva in podjetja, ki želijo ohraniti verodostojnost in preprečiti zlorabe. Vendar pa so detektorji le opozorilno orodje, ne gre za popolno znanost. Včasih sumljivo zaznajo celo besedilo, ki ga je napisal človek, in obratno. Njihovi algoritmi so zgolj statistični, ne forenzični, zato je ključno, da rezultate vedno preveri še človek.

Seveda obstajajo tudi načini, kako detektorje prelisičiti. Najpogostejši trik je ročno preurejanje stavkov, dodajanje človeških slovničnih napak, vključevanje osebnih zgodb ali uporaba "humanizerjev". To so posebna orodja, ki besedilo prepišejo v bolj "človeški" slog. Na spletu obstajajo celo vodiči, kako izkoristiti umetno inteligenco in hkrati skriti njen vir. Vendar takšne prakse odpirajo resna etična vprašanja: od šolske prevare do lažnega avtorstva v medijih ali celo zavajanja strank in bralcev. Odgovorna uporaba umetne inteligence pomeni, da avtor odkrito pove, kje in kako mu je orodje pomagalo. Tako umetna inteligenca ostaja pomočnik, ne skrivnost.

Tudi v Sloveniji se raziskovanje in uporaba naprednih tehnologij hitro razvija. Nacionalni kompetenčni center SLING v okviru razvojno-raziskovalnega projekta EuroCC2 nudi brezplačen dostop do superračunalniških zmogljivosti, podporo in izobraževanja uporabnikom. Superračunalniki v Mariboru (IZUM), Ljubljani (Arnes in IJS) ter Novi Gorici (Arctur) raziskovalcem, podjetjem in študentom omogočajo uporabo naprednih virov, hkrati pa pomagajo pridobiti dostop do še večjih sistemov evropske iniciative EuroHPC JU. Vse to odpira vrata, da bo v prihodnje tudi v Sloveniji vedno več ustvarjalcev vsebin uporabljalo lastne napredne modele in hkrati razvijalo detektorje za zaznavanje zlorab.

Praktično gledano lahko marsikaj prepoznamo tudi brez detektorja. Je besedilo nenavadno gladko in brez napak? Manjka oseben ton, konkretne izkušnje ali resnični primeri? So stavki splošni in se ponavljajo? To so lahko znaki, da je besedilo ustvarjeno z umetno inteligenco. Če sumimo na uporabo umetne inteligence, je detektor dober začetek, nikakor pa ne dokončen dokaz. Uredniki ali učitelji naj vedno zahtevajo osnutke, vire, razlage postopka dela in spodbudijo avtentičnost.

Zaznavanje besedil ustvarjenih z umetno inteligenco bo ostalo del vsakdanjika in še naprej bo šlo za igro med ustvarjalci vsebin in razvijalci detektorjev. A na koncu smo odgovorni mi: kako iskreno bomo priznali, kdo stoji za napisano besedo: človek, stroj ali oboje skupaj.



REPUBLIKA SLOVENIJA  
MINISTRSTVO ZA VISOKO ŠOLSTVO,  
ZNANOST IN INOVACIJE

SLING

EURO



Univerza v Mariboru

Projekt EuroCC 2 financira Evropsko Skupno podjetje za evropsko visokozmogljivo računalništvo (JU) v okviru sporazuma o dodelitvi sredstev št. 101101903. JU podpirajo EU program Digitalna Evropa, Nemčija, Bolgarija, Avstrija, Hrvaška, Ciper, Češka republika, Danska, Estonija, Finska, Grčija, Madžarska, Irska, Italija, Litva, Latvija, Poljska, Portugalska, Romunija, Slovenija, Španija, Švedska, Francija, Nizozemska, Belgija, Luksemburg, Slovaška, Norveška, Turčija, Republika Severna Makedonija, Islandija, Črna gora in Srbija. Delovanje Nacionalnega kompetenčnega centra SLING sofinancira RS, Ministrstvo za visoko šolstvo, znanost in inovacije.



EuroHPC  
Joint Undertaking



REPUBLIKA SLOVENIJA  
MINISTRSTVO ZA VISOKO ŠOLSTVO,  
ZNANOST IN INOVACIJE

Projekt EuroCC 2 financira Evropska unija. Financiran je s sredstvi Skupnega podjetja za evropsko visokozmogljivo računalništvo (EuroHPC JU) ter Nemčije, Bolgarije, Avstrije, Hrvaške, Cipra, Češke republike, Danske, Estonije, Finske, Grčije, Madžarske, Irske, Italije, Litve, Latvije, Poljske, Portugalske, Romunije, Slovenije, Španije, Švedske, Francije, Nizozemske, Belgije, Luksemburga, Slovaške, Norveške, Turčije, Republike Severne Makedonije, Islandije, Črne gore in Srbije v okviru sporazuma o dodelitvi sredstev št. 101101903.