

Nagy nyelvi modellek által generált szövegek felismerése magyar nyelven

Kiss Mihály¹, Berend Gábor¹

Szegedi Tudományegyetem, Informatikai Intézet
kiss.mihaly@stud.u-szeged.hu, berendg@inf.u-szeged.hu

Kivonat Manapság rengeteg mesterséges intelligencia (MI) által generált tartalom keletkezik, amit az is elősegít, hogy különböző modellek, mint például a ChatGPT, Claude vagy Gemini, széles körben elérhetők. Ebből adódóan egyre nagyobb szükség van olyan megoldásokra, amelyek képesek az ember és gép által generált tartalmak megkülönböztetésére. Míg angol nyelven már számos megoldás létezik, magyar nyelven egyelőre nem született igazán érdemi kutatás. A cikkben ezért kifejezetten magyar nyelvre készített és vizsgált modelleket mutatunk be az ember és MI által generált szövegek elkülönítésére.

Magyar nyelvre finomhangolt enkóder modelleket használtunk, öt különböző modell teljesítményét összehasonlítva. A modellek hasonló F1 értéket értek el. A szavaztatás a publikus korpuszon nem javított, a heterogénebb kiegészített korpuszon viszont növelte az F1-pontszámot.

Az eredményeket ipari detektorokkal is összevetettük, és azt találtuk, hogy a finomhangolt modellek több kereskedelmi megoldásnál is jobb teljesítményt nyújtanak, és csak egyetlen ipari detektor előzi meg őket.

Kulcsszavak: MI, detektálás, predikció, enkóder

1. Bevezetés

Az elmúlt években rengeteget fejlődtek a nagy nyelvi modellek, így az általuk generált szövegek sokkal jobban hasonlítanak az emberek által írt szövegekre, mint korábban. Manapság már rengeteg MI-generált szöveg keletkezik, és ez már a tudományos életben is észlelhető, egyre több publikációban, cikkben lehet szintetikus szöveget találni (Liang és mtsai, 2024b). Korábbi kutatások kimutatták, hogy az MI-konferenciákra beküldött cikkek 6,5–16,9%-a erősen támaszkodhat nagy nyelvi modellekre, vagy akár teljesen át is írhatták azokat, ami jól mutatja az egyre növekvő használatot a tudományos közegben (Liang és mtsai, 2024a). Ezt a trendet tovább erősíti, hogy a fejlett modellek széles körben elérhetők, és ma már minimális technikai tudással rendelkező felhasználók is hozzáférhetnek, használhatják azokat.

A fentiek miatt egyre nagyobb a szükség az olyan automatikus megoldásokra, amelyek képesek megkülönböztetni az ember által írt szövegeket a szintetikusan létrehozott szövegektől. Ezt az igényt számos közelmúltbeli verseny és workshop hangsúlyozza, amelyek kifejezetten erre a problémára készültek (Sarvazyan

és mtsai, 2023; Wang és mtsai, 2024; Chamezopoulos és mtsai, 2024; Wang és mtsai, 2025).

Noha egyre jobb minőségű megoldásokkal találkozhatunk, az MI segítségével létrehozott és az emberi szövegek megkülönböztetése sok kihívást rejt magában. A legjobb módszerek általában mély neurális hálózatu architektúrákra épülnek, legtöbbször transformer alapú modellek, amelyeket finomhangolnak nagyobb, annotált adathalmazokon (Sarvazyan és mtsai, 2024; Marchitan és mtsai, 2024). Frissebb tanulmányok megmutatták, hogy ezeknek a transformer modelleknek kombinálással vett valószínűségi kimenete még tovább javíthatja az eredményeket és a robusztusságot az MI-generált szövegek felismerésében (Gu és Meng, 2024; Kiss és Berend, 2025).

A területtel foglalkozó cikkek túlnyomó többségének fókuszában az angol nyelvre vonatkozó megoldások állnak. Jelen cikk elsődleges motivációja, hogy a magyar nyelven generált szövegek felismerésének problémáját – legjobb tudomásunk szerint – elsőként vizsgáljuk tudományos igényességgel, amely keretein belül egy új publikus tesztadalmat is létrehoztunk. A cikkben kettő megoldás eredményeit mutatjuk be. Az első megoldást három publikus adathalmazból összegyűjtött szövegekre, a másikat pedig egy korábban saját szkriptekkel gyűjtött adathalmazra támaszkodva hoztuk létre.

Az eredmények kiértékelése során fontos metrika a fals pozitív ráta (FPR) és a fals negatív ráta (FNR). Az FPR azt jelzi, hogy a modell milyen gyakran sorol tévesen emberi szöveget MI-generáltnak, míg az FNR azt mutatja meg, hogy az MI által generált szövegek mekkora részét nem ismeri fel. Ebben a doménben az egyik legfontosabb, hogy emberi szöveget ne osztályozzunk MI-generáltnak, ezért érdemes minél alacsonyabb fals pozitív rátára törekedni.

A két adathalmazra kapott eredmények azt mutatják, hogy a publikus adathalmazon finomhangolt modellek teljesítménye nem kiemelkedő, míg a saját gyűjtésű, kiegészített adathalmaz jóval robusztusabb és pontosabb modelleket eredményezett. A szavaztatásos módszerek a publikus adatokon nem hoztak érdemi javulást, a kiegészített adatokon azonban csökkentették a fals pozitív rátát és növelték az F1-pontszámot.

2. Kapcsolódó irodalom

A gép által generált szövegek detektálásában elért legújabb fejlesztések kiemelték a modellek szavaztatásával nyert (ensemble) tanulási módszerek hatékonyságát (Gu és Meng, 2024; Kiss és Berend, 2025). Gu és Meng (2024) finomhangolt transformer modellek szavaztatásával vett részt SemEval-2024 8. versenykiírásán (Wang és mtsai, 2024). Megközelítésük hatékonyan kezelte a kiegyensúlyozatlan címkék problémáját, amivel kimagasló teljesítményt voltak képesek elérni.

Mivel a standard ensemble módszerek magas számítási költséggel járnak, amelyek több független modell betanítását igénylik, számos megközelítést javasoltak a több modell döntésére építő megoldások erőforrásigényének csökkentésére. Az egyik ilyen módszer az ún. *snapshot ensemble* technika (Huang és mtsai, 2017), amely a finomhangolás különböző időpontjaiban rögzíti a modell aktuális

állapotának "pillanatfelvételeit". Bár ez a megközelítés csökkenti a finomhangolás költségeit azáltal, hogy elkerüli a különálló modellek betanításának szükségességét, általában nem éri el a standard ensemble módszer teljesítményét.

A felügyelt tanításon alapuló modellek mellett népszerű megközelítést jelentenek a gépileg generált szövegek felismerésének feladatában az ún. *zero-shot* módszerek (Mitchell és mtsai, 2023; Bao és mtsai, 2024). Ezen módszerek nem igényelnek finomhangolást, ehelyett (nagy) generatív nyelvi modelleknek az egyes szövegekhez társított valószínűségek elemzésére támaszkodnak.

3. Adathalmazok

3.1. Publikus adathalmaz

A publikus adathalmaz három másik adathalmazból lett készítve mintavételezéssel. Ezek az adathalmazok nyilvánosak, és elérhetőek. Az ember által írt szövegeket két forrásból szereztük be: egyrészt a magyar Wikipedia szócikkeiből Niklaus (2023), másrészt pedig a Hungarian Webcorpus 2 Nemeskey (2020) korpuszban található webes tartalmakból történő mintavételezésre támaszkodtunk. Ezáltal az adathalmazban reprezentálva vannak általános internetről kinyert szövegek, illetve tudományos szöveg is. Ebből a két adathalmazból lett véletlenszerűen kiválasztva 5000 szöveg, egyenlő arányban. A gép által generált szövegekre vonatkozó tanítóadat a boapps (2024) adathalmazból származik, amelyben magyar nyelven megfogalmazott utasításokra nagy nyelvi modellek által generált válaszok szerepelnek. Ebből az adathalmazból szintén kiválasztottunk 5000 szöveget, és így végső formában megkaptuk a 10,000 szövegből álló adathalmazunkat. A szövegek hosszának az elemzése a 2. ábrán látható. Egy rövid példa az MI generált szövegekből:

Kérdés: Írj egy rövid mondást az időről és annak múlásáról.

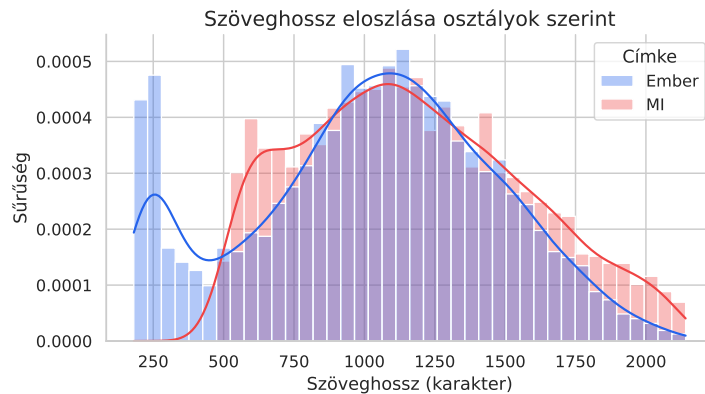
Válasz: Az idő olyan, mint a homok a tenyeredben – mindig csak úgy vesszük észre, hogy kicsúszik közöttünk.

1. ábra: Példa MI generált válaszra.

Alapvetően elmondható, hogy sok rövid szöveg is megtalálható az adathalmazban, amelyekről általában nehezen megmondható, hogy ki írta valójában.

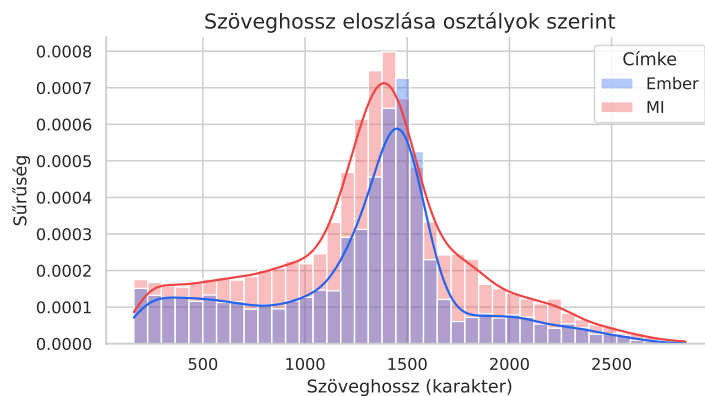
3.2. Kiegészített adathalmaz

A modellek fejlesztéséhez egy általunk összeállított kiegészített adathalmaz is felhasználásra került. A saját gyűjtésű adathalmazt kiegészítettnek nevezzük, mivel célja a publikus adathalmazban nem reprezentált szövegtípusok és nyelvi



2. ábra: Publikus tanítóhalmaz szövegek hosszának elemzése

jellemzők pótlása. Az adatgyűjtés és annotáció teljes mértékben a saját erőforrásainkból valósult meg. Az elmondható, hogy az adathalmaz diverz, sok témakört lefed. Ez alatt lehet érteni fórumok, esszék, tudományos cikkek, beadandókból előállított szövegeket. A kiegészített adathalmaz célja az volt, hogy a publikus adathalmazból hiányzó sokszínűséget lefedje, illetve az is, hogy megmutassuk azt, hogy egy minőségi adathalmazon finomhangolt modell jobb eredményt fog nyújtani. A szövegek hosszára vonatkozó statisztikák a 3. ábrán láthatók.

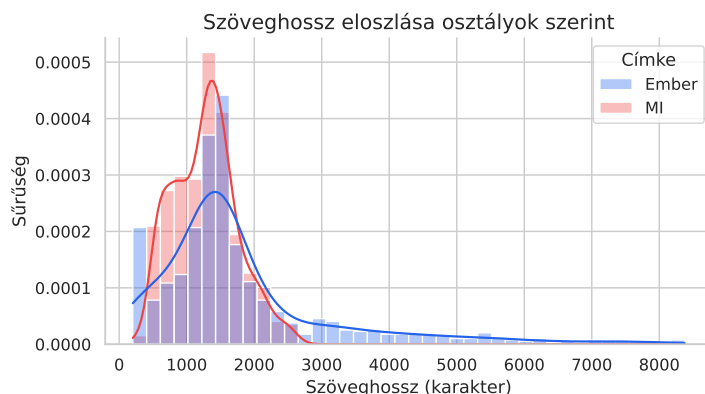


3. ábra: Kiegészített tanítóhalmaz szövegeinek hosszának elemzése

3.3. Teszt adathalmaz

A teszt adathalmaz a kiegészített illetve a publikus adathalmaz keveréke. Mindkettőből 1000-1000 szöveg lett beletéve, azon belül pedig 500-500 darab az em-

ber és az MI osztályból. Ezáltal a teszhalmaz diverz, amelyben mindkét kísérlet adathalmaza reprezentálva van, tehát mindkettő kísérletnek azonosan kedvez. A szövegek hossza a 4. ábrán látható. A teszt adathalmaz nyilvánosan elérhető¹.



4. ábra: Teszhalmaz szövegek hosszának elemzése

4. Felhasznált modellek

A kísérletekben több, magyar nyelvre finomhangolt transformer alapú modell nyelvi modellt használtunk. Ezáltal nem csak az adathalmazok minőségét tudjuk vizsgálni, hanem a modellek különböző előtanítási megoldásait is. A huBERT (Nemeskey, 2021), PULI-BERT-Large (Yang és mtsai, 2023) és a huDeBERTa-MLM modellek klasszikus, maszkolt nyelvi modellezésen (MLM) alapuló előtanítással készültek, amely a BERT-jellegű modellek általános előtanítási feladata. A huDeBERTa-MLSM illetve a huDeBERTa-LSM a standard maszkolt nyelvi modellezéses előtanítási feladat helyett látens szemantikus tulajdonságok modellezésére lettek tanítva a Berend (2024) által bevezetett megoldások szerint.

A huDeBERTa prefixű modellek azzal az előnyös tulajdonsággal rendelkeznek, hogy azok az előtanítás célfüggvényét leszámítva teljesen azonos módon lettek létrehozva (azonos modellarchitektúra, azonos előtanító korpusz, illetve azonos előtanítási hiperparaméterek használata mellett), így az ezen modellek finomhangolásával nyert eredmények összehasonlítása lehetővé teszi az előtanítás megválasztásának szerepének vizsgálatát.

A vizsgált modellek méreteiben jelentős különbségek vannak: a PULI-BERT-Large egy 24 rétegű modell 340 millió paraméterrel, míg a huBERT-base és a huDeBERTa család modelljei 12 rétegből állnak, előbbi 110, utóbbi pedig 134 millió paraméterrel. A különböző modellek használata lehetővé teszi annak vizsgálatát

¹ https://huggingface.co/datasets/mihalykiss/Hun_ai_test

is, hogy a modellek méretei milyen mértékben járulnak hozzá az MI-generált szövegek felismerésének pontosságához.

5. Kísérletek

5.1. Modellek finomhangolása

Az előző pontban bemutatott modelleket mindkét adathalmaz alapján finomhangoltuk. A finomhangolások során alkalmazott hiperparamétereket a tanítóadattól és az aktuálisan finomhangolt modelttől függetlenül, azonos módon választottuk meg a kapott eredmények összehasonlíthatósága kedvéért. Az adatbázisokat minden esetben azonos módon – 80% tanító, 20% validációs halmaz – osztottuk fel.

A finomhangolás során minden epoch végén elvégeztünk egy validációs halmazon történő kiértékelést. Alapértelmezetten modellek finomhangolása 5 epochon keresztül zajlott, de lehetőség volt az ennél korábban történő megállásra (early stopping) a validációs halmazon elért F1-pontszám figyelembe vétele alapján. Minden alapmodellel 3 finomhangolási kísérletet hajtottunk végre úgy, hogy a modelleknek az osztályozásért felelős súlyait különböző véletlenszerű inicializációval hoztuk létre. Egy finomhangoláson belül a validációs halmazon legjobb eredményt elérő modellt tartottuk meg.

Mindegyik vizsgált enkóder alapmodell rendelkezik egy 512 tokenes bemeneti korláttal, amennyiben egy input szövegnek a hossza ezt meghaladta, úgy az utána következő részt levágtuk. A kísérletek során alkalmazott további hiperparamétereink a következők voltak: tanulási ráta: $2e-5$, lineáris tanulási ráta csökkentés, warmup ratio = 0,1.

5.2. Szavaztatásos megoldások

A tradicionális szavaztatásos megoldások, – ahol több egymástól függetlenül tanított modell együttese hozza meg a végső predikciót – széles körben ismert és használt megoldás, amellyel az egyes modellek eredményessége és robusztussága jellemzően javítható. Ez a megoldás természetesen a mesterséges intelligencia által generált szövegek felismerésére is alkalmas (Abburi és mtsai, 2023; Gu és Meng, 2024).

A megoldásainkban szavaztatásos megközelítést alkalmaztunk, amely során az azonos alapmodell különböző véletlenszerű inicializációval finomhangolt változatainak valószínűségi kimeneteit uniform módon átlagoltuk a végső predikció meghatározásához. Általában ez a megoldás javulást szokott hozni az eredmények tekintetében, azonban az erőforrásigénye drasztikusan megnő, mivel több modellt kell finomhangolni, és a kiértékelés költsége is az alkalmazott modellek számával arányosan nő.

6. Eredmények

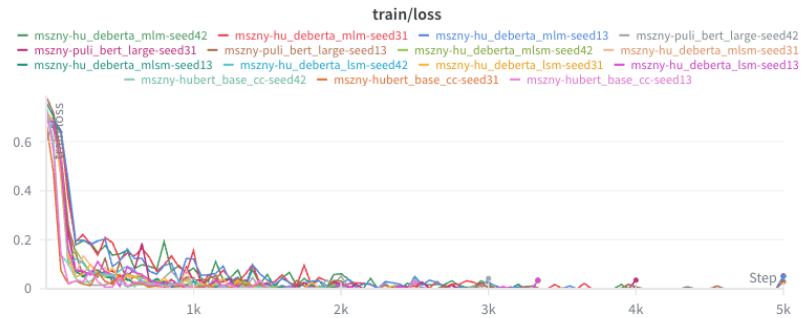
6.1. Publikus adathalmaz eredményei

A publikus adathalmazon történő finomhangolás eredményei azt mutatják az 1. táblázatban, hogy a modellek nagyon közeli eredményeket produkálnak. Ettől valamelyest lemaradnak a huDeBERTa-MLM modell finomhangolásával kapott eredmények.

A legjobb átlagos eredményt a PULI-BERT-Large modell érte el, ugyanakkor megjegyzendő, hogy ennek a modellnek a kapacitása a többi vizsgált modell kapacitásának többszöröse, így valós gyakorlatban való alkalmazása a többi finomhangolt modell használatához képest számottevő erőforrásigénnyel rendelkezik (egy 24 rétegből álló, 1024 dimenziós rejtett reprezentációkat alkalmazó modelltől van szó, míg a többi vizsgált modell csupán 12 rétegből áll, és a rejtett vektoraik 768 dimenziósak).

1. táblázat. Publikus adathalmazon tanított modellek eredményeinek átlaga

Modellek	Pontosság	F1-pontszám	FPR	FNR
huBERT-base	$0,786 \pm 0,009$	$0,754 \pm 0,013$	$0,082 \pm 0,008$	$0,345 \pm 0,017$
PULI-BERT-Large	$0,794 \pm 0,002$	$0,770 \pm 0,005$	$0,100 \pm 0,017$	$0,311 \pm 0,016$
huDeBERTa-MLM	$0,759 \pm 0,005$	$0,700 \pm 0,011$	$0,046 \pm 0,013$	$0,436 \pm 0,021$
huDeBERTa-MLSM	$0,789 \pm 0,011$	$0,759 \pm 0,016$	$0,085 \pm 0,002$	$0,337 \pm 0,023$
huDeBERTa-LSM	$0,782 \pm 0,005$	$0,746 \pm 0,013$	$0,074 \pm 0,016$	$0,361 \pm 0,027$



5. ábra: Publikus adathalmazon tanított modellek tanítási veszteségének ábrázolása

Az 5. ábrán jól megfigyelhető, hogy a tanulási folyamat az elején meredeken, gyorsan konvergál, és a veszteség már a korai fázisokban is alacsony. Ez egy

általános probléma ebben a témakörben, amely azért gond, mivel a modell nem tanul lényeges új mintázatokat.

A szavaztatásos módszerek ebben az esetben nem hoztak érdemi javulást. Ez arra utalhat, hogy a publikus korpusz nem tükrözi a valóságban fellelhető magyar nyelvű szövegek heterogenitását, és nem tartalmaz elég nehezen osztályozható példát. A szavaztatásos módszerek eredménye a publikus adathalmazra a 2. táblázatban található. A fals pozitív rátán (FPR oszlop) a legtöbb esetben segített a szavaztatás, amely egy kritikus érték ebben az osztályozási feladatban.

2. táblázat. Publikus adathalmaz modelljei szavaztatva

Modellek	Pontosság	F1-pontszám	FPR	FNR
huBERT-base szavaztatás	0,786	0,752	0,078	0,350
PULI-BERT-Large szavaztatás	0,793	0,767	0,094	0,320
huDeBERTa-MLM szavaztatás	0,760	0,703	0,050	0,431
huDeBERTa-MLSM szavaztatás	0,792	0,762	0,082	0,334
huDeBERTa-LSM szavaztatás	0,780	0,741	0,072	0,369

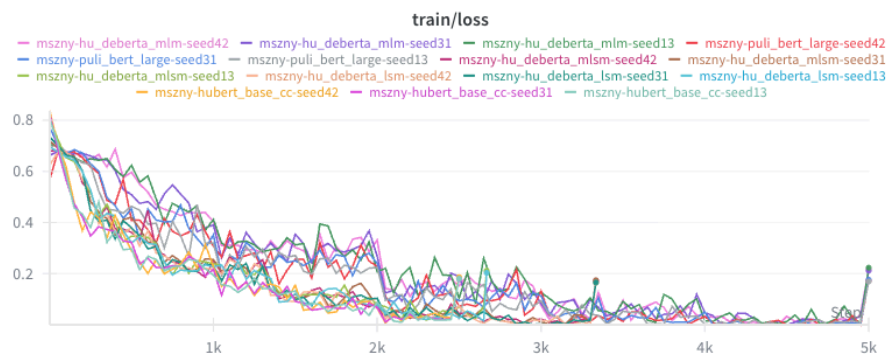
6.2. Kiegészített adathalmaz eredményei

A kiegészített adathalmazon kapott eredmények jelentősen felülmúlják a publikus adatok használatával mért értékeket. A legjobb eredményt önálló modellként a huDeBERTa-MLSM érte el 0,905 F1-pontszámmal, a további eredmények a 3. táblázatban láthatóak.

3. táblázat. Kiegészített adathalmazon tanított modellek eredményeinek átlaga

Modellek	Pontosság	F1-pontszám	FPR	FNR
huBERT-base	$0,885 \pm 0,021$	$0,891 \pm 0,015$	$0,165 \pm 0,071$	$0,065 \pm 0,030$
PULI-BERT-Large	$0,899 \pm 0,006$	$0,902 \pm 0,003$	$0,138 \pm 0,035$	$0,065 \pm 0,022$
huDeBERTa-MLM	$0,827 \pm 0,015$	$0,841 \pm 0,012$	$0,257 \pm 0,033$	$0,089 \pm 0,013$
huDeBERTa-MLSM	$0,901 \pm 0,005$	$0,905 \pm 0,003$	$0,139 \pm 0,033$	$0,059 \pm 0,023$
huDeBERTa-LSM	$0,883 \pm 0,005$	$0,888 \pm 0,008$	$0,164 \pm 0,023$	$0,070 \pm 0,032$

A veszteségfüggvények viselkedése a kiegészített tanítóhalmazon történő finomhangolás esetén jelentősen eltér a korábban látottaktól, ahogy azt az 5 és a 6. ábrák összevetése is jól mutatja. Míg a korábbi esetben látott veszteségfüggvény-értékek nagyon hamar 0 környékévé váltak, addig ebben az esetben a veszteségfüggvény csak később kezd el csökkenni, ami azt jelzi számunkra, hogy a modell sokkal komplexebb példákat kapott a kiegészített adatokon történő tanítás során.



6. ábra: Kiegészített adathalmazon tanított modellek tanítási veszteségének ábrázolása

Míg a publikus adathalmazon a szavaztatás hatása nem volt jelentős, a kiegészített adathalmazon történő finomhangolás során minden modell esetében javulást eredményezett az F1-pontszám tekintetében, ahogy azt a 4. táblázatban láthatjuk. A legjobb teljesítményt nyújtó modellnek a szavaztatást követően is a huDeBERTa-MLSM modell bizonyult, csakúgy mint az önálló eredményeknél. Az is látható, hogy a szavaztatás minden esetben képes volt a fals pozitív ráta csökkentésére. A kapott eredmények alapján a kiegészített adathalmaz sokkal inkább tekinthető valós felhasználásra alkalmasnak, mivel a magyar nyelvű szövegek jellemzőit jobban reprezentálja, amelyet a modell is jobban meg tud tanulni.

4. táblázat. Kiegészített adathalmaz modelljei szavaztatva

Szavaztatott modellek	Pontosság	F1-pontszám	FPR	FNR
huBERT-base	0,892	0,898	0,164	0,052
PULI-BERT-Large	0,909	0,912	0,128	0,054
huDeBERTa-MLM	0,834	0,846	0,247	0,085
huDeBERTa-MLSM	0,912	0,916	0,130	0,046
huDeBERTa-LSM	0,895	0,900	0,153	0,057

6.3. Külső (kereskedelmi) detektor eredményei

A saját finomhangolt megoldásainkat összevetettük több piacon elérhető detektorral². A tesztelt kereskedelmi detektorok mindegyike támogatja a magyar

² <https://preds.hu>, <https://pangram.com>, <https://zerogpt.com>

nyelvet. API hívásokon keresztül küldtünk ki kéréseket, a korábban bemutatott tesztalmazra. Az 5. táblázatban láthatóak az eredmények, amelyről elmondható, hogy a kereskedelmi detektorok közül egyedül a Preds modellje ért el az általunk bemutatottaknál jobb eredményt, és ennek a modellnek a legalacsonyabb a fals pozitív rátája is.

5. táblázat. Modellek összehasonlítása

Modellek	Pontosság	F1-pontszám	FPR	FNR
Legjobb publikus modell	0,793	0,767	0,094	0,320
Legjobb kiegészített modell	0,912	0,916	0,130	0,046
Preds	0,954	0,952	0,007	0,085
PangramLabs	0,789	0,744	0,034	0,388
ZeroGPT	0,554	0,509	0,352	0,539

7. Összegzés

A célunk az volt, hogy tudományos igényességgel elsőként vizsgáljuk az MI-generált szövegek felismerésének lehetőségeit. Több adathalmazzal kísérleteztünk, az első adathalmaz egy publikus, már kész adathalmazokból összeszedett korpusz, a második pedig egy általunk gyűjtött adathalmaz, amelyben diverz forrásokból szerzett szövegek vannak.

A publikus korpuszon a modellek gyorsan, már az első epochokban 0 közeli veszteséget mutatnak, ami arra utal, hogy a feladat túl könnyű volt. A legjobb publikus eredmény szavaztatás nélkül 0,770 F1-pontszámot ért el, szavaztatással pedig 0,767 F1-pontszám volt, ahol a szavaztatás általában nem hozott javulást.

Ezzel szemben a kiegészített adathalmaz heterogénebb adathalmaz lett, ahol lényegesen jobb eredmények születtek. A legjobb önálló modell 0,905 F1-pontszámot (huDeBERTa-MLSM), míg a szavaztatásos megoldás 0,916 F1-pontszámot ért el. Ez azzal állhat összefüggésben, hogy a kiegészített adathalmaznál a modellek predikciói diverzebbek voltak, amely pedig az adathalmaz nagyobb variációjából adódhatott.

A megoldásainkat összevetettük több piacon elérhető detektorral is, ahol a Preds modellje érte el a legjobb teljesítményt a pontosság és az F1-pontszám tekintetében. A másik két kereskedelmi detektor gyengébb eredményeket mutatott, mint az általunk létrehozott publikus modellek. A Preds modellje kiemelkedően alacsony fals pozitív rátát produkált.

Ennek egyik lehetséges oka, hogy a kereskedelmi megoldások várhatóan jóval nagyobb és diverzebb tanítóadaton voltak betanítva, illetve a problémára optimalizált hiperparaméterekkel és megoldásokkal finomhangolták a modelleket. Eredményeink ugyanakkor azt mutatják, hogy már a publikus, kisebb adathalmazokon finomhangolt modellekkel is elérhető némely kereskedelmi detektor pontossága.

Köszönetnyilvánítás

A cikk a Bolyai János Kutatási Ösztöndíj támogatásával készült. A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A MILAB SZTE nyelvtechnológia projekt nevében köszönetet mondunk az ELKH Cloud (lásd: Héder és mtsai (2022); <https://science-cloud.hu/> használatáért, ami hozzájárult a publikált eredmények eléréséhez.

Hivatkozások

- Abburi, H., Roy, K., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E., Bhattacharya, S.: A simple yet efficient ensemble approach for AI-generated text detection (2023), <https://arxiv.org/abs/2311.03084>
- Bao, G., Zhao, Y., Teng, Z., Yang, L., Zhang, Y.: Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=Bpcgcr8E8Z>
- Berend, G.: Neurális nyelvi modellek látens szemantikus információ alapján történő maszkolásmentes előtanítása. In: XX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 85–95. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2024), <https://acta.bibl.u-szeged.hu/78418/>
- boapps: alpaca-hu-v2: Hungarian instruction-follower dataset. <https://huggingface.co/datasets/boapps/alpaca-hu-v2> (2024), accessed: 2025-11-22
- Chamezopoulos, S., Herrmannova, D., De Waard, A., Herrmannova, D., Rosati, D., Kashnitsky, Y.: Overview of the DagPap24 shared task on detecting automatically generated scientific paper. In: Ghosal, T., Singh, A., Waard, A., Mayr, P., Naik, A., Weller, O., Lee, Y., Shen, S., Qin, Y. (szerk.) Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024). pp. 7–11. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), <https://aclanthology.org/2024.sdp-1.2>
- Gu, R., Meng, X.: AISPACE at SemEval-2024 task 8: A class-balanced soft-voting system for detecting multi-generator machine-generated text. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (szerk.) Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). pp. 1476–1481. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), <https://aclanthology.org/2024.semeval-1.212>
- Héder, M., Rigó, E., Medgyesi, D., Lovas, R., Tenczer, S., Török, F., Farkas, A., Emődi, M., Kadlecsek, J., Mező, Gy., Pintér, Á., Kacsuk, P.: The past, present and future of the ELKH cloud. Információs Társadalom 22(2), 128 (aug 2022), <https://doi.org/10.22503/inftars.xxii.2022.2.8>
- Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free. In: International Conference on Learning Representations (2017), <https://openreview.net/forum?id=BJYwwY911>

- Kiss, M., Berend, G.: SzegedAI at GenAI detection task 1: Beyond binary - soft-voting multi-class classification for binary machine-generated text detection across diverse language models. In: Alam, F., Nakov, P., Habash, N., Gurevych, I., Chowdhury, S., Shelmanov, A., Wang, Y., Artemova, E., Kutlu, M., Mikros, G. (szerk.) Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect). pp. 166–172. International Conference on Computational Linguistics, Abu Dhabi, UAE (Jan 2025), <https://aclanthology.org/2025.genaidetect-1.15/>
- Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., Mcfarland, D., Zou, J.Y.: Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (szerk.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 29575–29620. PMLR (21–27 Jul 2024a), <https://proceedings.mlr.press/v235/liang24b.html>
- Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C.D., Zou, J.Y.: Mapping the increasing use of LLMs in scientific papers. In: First Conference on Language Modeling (2024b), <https://openreview.net/forum?id=YX7QnhxESU>
- Marchitan, T.g., Creanga, C., Dinu, L.P.: Team Unibuc - NLP at SemEval-2024 task 8: Transformer and hybrid deep learning based models for machine-generated text detection. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (szerk.) Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). pp. 403–411. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), <https://aclanthology.org/2024.semeval-1.63/>
- Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. ICML’23, JMLR.org (2023)
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020)
- Nemeskey, D.M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021). pp. 3–14. Szeged (2021)
- Niklaus, J.: EU Wikipedias Dataset. https://huggingface.co/datasets/joelniklaus/EU_Wikipedias (2023), accessed: 2025-11-22
- Sarvazyán, A.M., González, J.Á., Franco-salvador, M.: Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (szerk.) Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). pp. 101–107. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), <https://aclanthology.org/2024.semeval-1.17>
- Sarvazyán, A.M., González, J.Á., Franco-Salvador, M., Rangel, F., Chulvi, B., Rosso, P.: Overview of AuTextification at IberLEF 2023: Detection and att-

- tribution of machine-generated text in multiple domains. *Proces. del Leng. Natural* 71, 275–288 (2023), <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6559>
- Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Mohammed Afzal, O., Mahmoud, T., Puccetti, G., Arnold, T.: SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (szerk.) *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. pp. 2057–2079. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), <https://aclanthology.org/2024.semeval-1.279>
- Wang, Y., Shelmanov, A., Mansurov, J., Tsvigun, A., Mikhailov, V., Xing, R., Xie, Z., Geng, J., Puccetti, G., Artemova, E., Su, J., Ta, M.N., Abassy, M., Elozeiri, K., Ahmed, S.E.D., Goloburda, M., Mahmoud, T., Tomar, R.V., Aziz, A., Laiyk, N., Afzal, O.M., Koike, R., Kaneko, M., Aji, A.F., Habash, N., Gurevych, I., Nakov, P.: GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In: *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*. International Conference on Computational Linguistics, Abu Dhabi, UAE (January 2025)
- Yang, Z.Gy., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, A., Laki, L.J., Ligeti-Nagy, N., Vadász, N., Váradi, T.: Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In: *XIX. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 247–262. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2023), <https://acta.bibl.u-szeged.hu/78417/>