TDK Dolgozat

Fodor Gábor, Sánta Péter

Gépelted vagy GÉPelted? ChatGPT és Gemini leleplezése magyar szövegekben

Fodor Gábor, III. évf. programtervező informatikus BSc
Sánta Péter, II. évf. programtervező informatikus MSc
Témavezetők: Dr. Kicsi András, Dr. Vidács László
Szegedi Tudományegyetem, Természettudományi és Informatikai Kar,
Szoftverfejlesztés tanszék

Tartalomjegyzék

Absztrakt

Az elmúlt években berobbantak a nagy nyelvi modellek a hétköznapi életbe is, egyre elterjedtebb a használatuk számtalan felhasználási körben. Az OpenAI ChatGPT-jének 2022-es megjelenése óta világszerte íratják meg mesterséges intelligencia segítségével tanulók a házi feladataikat, illetve esszéiket; egyetemi hallgatók a beadandóikat; különböző híroldalak a cikkeiket. A Google Gemini modelljének 2023-as megjelenésével pedig még több eszköz áll az emberek rendelkezésére szövegek gépi generálásához. Problémákkal is járnak azonban ezek a modellek, mivel szövegeikről gyakran kihívás megállapítani, hogy nem ember írta őket, ezzel pedig súlyosan félre lehet vezetni embereket.

Kutatásunk során ezen probléma megoldásán dolgoztunk, olyan gépi tanuló modellek tanításával foglalkoztunk, melyek képesek megkülönböztetni az ember, a ChatGPT 4.0 és a Gemini által írt szövegeket, 3 különböző osztályba sorolva őket. Ilyen módon megállapítható, hogy megmondható-e, hogy egy adott szöveget ember írta-e, ahogy az is, hogy kétféle nyelvi modell által generált szövegben kimutathatók-e jelentős különbségek. A használt modell további nyomokat adhat a vizsgálódó személyeknek egy csalás, vagy súlyosabb esetekben akár egy bűncselekmény tekintetében: például, ha az illető fenyegetőlevelek gyártására használta a modellt annak érdekében, hogy megpróbálja elrejteni nyomait.

Kísérleteink alatt elsősorban magyar, hétköznapi nyelvezetű szövegekre fókuszáltunk. A gépi szövegek generálásához különböző, véletlenszerűen generált személyiségeket alkalmaztunk a promptolás során annak érdekében, hogy több különböző stílusú gépi szövegből álljon adathalmazunk. Az emberi szövegek különböző szerzőktől származtak ugyanezen okból. Több különböző módon válogatott, feldolgozott adathalmazzal is kísérleteztünk, többek között vizsgáltuk mennyire befolyásolja a modell teljesítményét az elírások száma, a halmozott írásjelek használata, illetve a szöveg hossza. Modelljeink konzisztensen magas pontossággal teljesítettek a különböző adathalmazokon, és bár bizonyos utófeldolgozási lépések a gépi kimenetre csökkentették a teljesítményt, az ilyen adathalmazokon is pontos maradt a modell.

Módszertanunkat a HuBERT nyelvi modellre alapoztuk, illetve vizsgálatokat

végeztünk azt illetően, hogy különböző nyelvi jellemzők hogyan befolyásolják a modell pontosságát. Ezen jellemzőket egy saját készítésű, főként a HuSpaCy nyelvi elemzésére építő elemzővel nyertük ki. A kinyert nyelvi jellemzőket a Multimodal Toolkit programcsomag segítségével kombináltuk különböző módszerekkel a HuBERT által adott jellemzővektorokkal.

1. Bevezetés

A huszonegyedik század rengeteg tudományos, technológiai áttöréssel, fejlődéssel rendelkezik. Ezek közé tartozik a mesterséges intelligencia rohamos feltörekvése is. Mára már rengetegen dolgoznak olyan technológiával, mely hátterében dolgozik legalább egy neurális háló modell, a ChatGPT [29] 2022-es köztudatba való betörésének köszönhetően pedig a mesterséges intelligencia a hétköznapi felhasználáshoz is minden eddiginél közelebb került. A generatív mesterséges intelligencia berobbanása óta magas fokú népszerűségnek örvend; rengeteg kép-, szöveg-, videó-, és még számtalan digitális anyag készül neurális hálók segítségével. A rengeteg generált anyaggal kapcsolatban azonban rengeteg probléma felmerülhet, többek között; plagizált munkának minősülhetnek, obszcén, nem kívánatos tartalommal rendelkezhetnek, nehéz lehet megállapítani, hogy tényleg mesterséges intelligencia alkotta-e, vagy sem. Kutatásunk az utóbbi problémára fókuszált: megállapítható-e maximális bizonyossággal egy adott szövegről, hogy az egy ember, vagy neurális háló "tollából" származik [32]? A kérdés több másikat is felvethet: tud-e egyáltalán konzisztensen jó, meggyőző szövegeket írni egy neurális háló? Ha tud, akkor miként tudjuk módosítani a kimeneti szöveg hosszát, minőségét, témáját, töredezettségét? Lehetséges-e olyan promptot (bemeneti adatot – esetünkben szöveget –, mely alapján a modell tartalmat generál) létrehozni, mellyel teljesen személyre lehet szabni a megkapni kívánt szöveget? A válaszok megadásához a természetesnyelv feldolgozás [[4], [23]] és nagy nyelvi modellek eszközeit alkalmaztuk. A természetesnyelv feldolgozás (natural language processing - NLP) a mesterséges intelligencia azon ága, amely az emberi nyelvek számítógépek általi feldolgozásával foglalkozik. A ma is használatos technológiák (mint például szöveg- elemzés és alkotás, internetes fordítás, információ kinyerés, chatbotok) is a természetesnyelv feldolgozáson alapulnak. A nagy nyelvi modellek azok a nagy szövegmennyiségen tanított neurális hálók, melyekkel NLP alkalmazása során a legtöbbször dolgozunk. Célunk egy olyan módszer meghatározása, mely magyar nyelvű szövegek szétválasztására képes: ember, vagy egy generatív mesterséges intelligencia írta-e, ha pedig utóbbi a szerző, akkor pontosan melyik neurális háló modell áll a háttérben.

Munkánk során három különböző osztályba soroltunk hétköznapi szövegeket: ember által írt, ChatGPT 4.0 [25] által íródott, illetve a Google Gemini 1.0 Pro modellje [33] által íródott szövegek. A ChatGPT 4.0 esetén a 2024 február 13-as kiadás alapján dolgoztunk, míg a Gemini 1.0 Pro esetén a 2024 április 9-es kiadás alapján. A módszertanhoz a saját készítésű nyelvi elemzőnk [15] továbbfejlesztett változatát, mely mögött a HuSpaCy [26], Hunspell 1.7.0 [20], illetve Benepar [[18], [17]] modellek állnak, valamint a HuBERT [24] nagy nyelvi modellt használtuk fel a Multimodal Toolkit [9] (multimodális jellemző-kombinálási csomag) segítségével.

2. Háttér

2.1. A generatív mesterséges intelligencia és nagy nyelvi modellek működési elve

Miként működik egy neurális háló, mi történik belül? A legtöbb ma is használatos mesterséges intelligencia azok felépítéséből adódóan egy-egy fekete doboz modellként viselkednek; nem tudjuk, hogy miként, miket tanult meg az adott modell, milyen összefüggéseket talált meg, és azok helytállóak-e. Egy ajánlórendszerből nem tudjuk visszafejteni, hogy az egyes ajánlásokat milyen kritériumok teljesülése alapján tette a felhasználó felé, csupán értékelni tudjuk, hogy azok mennyire voltak jók, relevánsak. A technológia népszerűségének köszönhetően rengeteg kutatás irányul ezen fekete doboz működési elv teljeskörű megértésére [10] és feloldására, valamint az (emberileg) értelmezhető mesterséges intelligenciára [11]. Mivel munkánk során két nagy nyelvi modellt felhasználó generatív mesterséges intelligenciával dolgozunk, ezért érdemes bemutatni a technológiát: a generatív mesterséges intelligencia egy olyan mesterséges intelligencia, mely több különböző jellegű, típusú adaton lett tanítva – többek között szövegeken, képeken, programkódokon, akár 3D modelleken –, a felhasználó pedig különböző típusú bemenetet – a tanult adat típusai közül – megadva szintén bármilyen, a háló értelmezésében megfelelő típusú kimenetet tud generálni. Például egy generatív mesterséges intelligencia modell egyszerre lehet képes szövegből kép (text to image), illetve képből szöveg (image to text)

előállítására. Gyakorlati szempontból egy generatív mesterséges intelligencia sokkal több feladatot tud elvégezni, mint az "egyszerűbb" társai, ez azonban a teljesítmény kárára történhet; egy generatív mesterséges intelligencia általában egyik feladatát sem képes olyan hatékonysággal elvégezni, mint a külön-külön, feladatspecifikusan tanított neurális hálók többsége; más szóval kevés esetben várható el tőlük a legkorszerűbb (state-of-the-art) modellekhez hasonló teljesítmény.

2.2. Adatok generálása a ChatGPT és Gemini által

A neurális hálók működési elvéből kifolyólag kutatómunkánk első mérföldköve a ChatGPT és Google Gemini működésének megértése gyakorlati szempontból: nem célunk pontosan meghatározni, milyen aktivációs függvények lépnek érvénybe egy-egy folytonos szózsák generálása során, vagy mennyire értelmezhetőek a modellek működésüket, felépítésüket tekintve, csupán azt kell megállapítanunk, milyen típusú promptra milyen jellegű szöveget ad vissza az adott chatbot; ezt a folyamatot, melynek keretén belül meg tudjuk tervezni a leghatékonyabb promptot (kérdést) a nyelvi modellnek, nevezzük prompt engineering-nek [7]. A megfelelő promptolási/kérdésfeltevési tudás hiányában kihívást jelenthet megfelelő minőségű szöveget generáltatni; a ChatGPT-t megjelenésekor sok kritika érte a kimenetei minőségét illetően – ezen kritikák azonban magas százalékban olyan felhasználási formákból érkeztek, melyek esetében az adott felhasználó nem rendelkezett elegendő belátással a használt nyelvi modell működését [16] illetően.
A szóban forgó chatbotok teljesítményét, minőségét nem csak a neurális hálók általános működési elvei határozzák meg, hanem a fejlesztők által kézzel, vagy egyéb modellekkel belekódolt működési szabályok. A bemeneti és/vagy kimeneti szöveg hosszát tekintve rendelkezik felső korláttal, illetve trágárságra, valamint illegális tartalmak gyártására nagyon nehéz kötelezni a chatbotokat. A tudomány és a technológia mai állása szerint a jelenlegi neurális hálók sérülékenyek; megfelelően manipulált adat bemenetként való megadása a neurális háló összezavarodásához vezet, ami hibás, vagy akár teljesen rossz, elfogadhatatlan kimenetet eredményez. Képfelismerés esetén a háló egy telje-

sen más osztályba sorolja a besorolandó képet (például kutya helyett pandát prediktál), szöveggenerálás esetén pedig halandzsa szöveg generálódik. Ezeket a célzottan hálók összezavarására szánt bemeneteket nevezik ellenséges példáknak [38] (adversarial examples); maga a folyamat az ellenséges támadás (adversarial attack) nevet viseli. Az ellenséges példák több, különböző módon fejthetik ki a hatásukat: támadhatják az egész hálót, annak csupán egy adott (például a kimeneti) rétegét, megfelelő információs tér esetében a teljes tanuló algoritmus logikáját is felülírhatja [2] a példa. A támadás jellege, erőssége a neurális háló felépítésétől is nagyban függ. Az ellenséges példák ellen a szakirodalom robusztus neurális hálók alkalmazását javasolja [21], azonban kutatók már azt is kimutatták, hogy a robusztusnak vélt hálók nem feltétlenül robusztusak [12], ebből kifolyólag pedig nem megfelelőek a támadások elleni védekezésre.

Tay, a Microsoft 2016-ban napvilágot látott chatbotja az aktiválását követően kevesebb, mint 16 órán belül leállításra [30] került. Az előzetes tervek szerint az akkor még Twitternek nevezett felületen kommunikált volna a felhasználókkal. Különlegessége volt, hogy előtanított mesterséges intelligencia volt, azonban a felhasználókkal való interakciók során is folyamatosan tanult. A mesterséges intelligencia felhasználó-interaktív tanulása egy ellenőrizetlen folyamat volt, melynek eredményeképp az hamar rasszista, gyűlölködő, fenyegető üzeneteket kezdett el megosztani a platformon. Tay esetéből jól látható, szükség van a chatbotok szabályozására. Tapasztalatok alapján sem a Gemini, sem pedig a ChatGPT 4.0 nem tanul közvetlenül a felhasználóktól – azonban tapasztalatok alapján feldolgozott, tényszerű adatokkal a hálók bizonyos időközönként tanításon esnek át. A két chatbot a látszólagos folyamatos tanítás, illetve szabályok mellett is támadhatóak; ChatGPT-vel sikeresen generáltattak Windows 10 generikus licenc kulcsokat [3], melyek részleges, illetve Windows 95 licenc kulcsokat [36], melyek teljes hozzáférést nyújtottak az adott operációs rendszer funkcióihoz. Az ilyen szövegkimenetek eléréséhez leggyakrabban egy elképzelt szituáció, vagy egy erős érzelmi alapú indok szükséges a promptba. Ezen felül fontos belátni, hogy a szabályozások, kézzel bekódolt szabályok sem mindig működnek elsőre – a Gemini megjelenésekor rendelkezett egy képgeneráló funkcióval, melybe a fejlesztők belekódoltak egy antirasszista

szabályt; a Gemini által kigenerált képeken az ábrázolt személyek nem tartozhatnak minden esetben egyazon etnikumhoz. A gyakorlatban a neurális háló a szabályozás miatt látszólag teljesen figyelmen kívül hagyta az etnikai és történelmi pontosságot a promptok esetében, így generálva potenciálisan sértő képeket [31].
Mesterséges intelligencia által generált szövegek felismerésére több módszer is létezik, szabály alapú megközelítés, vagy akár egy erre a feladatra tanított neurális háló is megfelelő eszköz lehet; kutatásunk során mi az utóbbi megközelítésre fókuszáltunk. Általános esetben már igazolt, hogy lehetséges a neurális hálók által megírt szöveg pontos detektálása [37] (akár nem csak teljes szövegekre, hanem mesterséges intelligencia által íródott szövegtöredékekre is [8]). Munkánk során a legkorszerűbb eredmények reprodukálását kíséreljük, teljesen magyar szövegek tekintetében: Emellett azt is megvizsgáljuk, lehetséges-e utófeldolgozással olyan ellenséges példákat létrehozni a generált, illetve emberileg írott szövegek tekintetében, melyekkel a detektáló hálónkat képesek lehetünk összezavarni, ez által rontani annak teljesítményén.
Az emberi és ChatGPT 4.0 adathalmazok kialakítása során kutatásunk kevés akadályba ütközött. A Gemini adathalmaz esetén azonban jelentős akadálynak bizonyult a szövegek legenerálása azon okból, hogy a Gemini API nem elérhető az Európai Unió területén, így Magyarországon sem. [1] Ennek következtében szükségesnek bizonyult kézzel legeneráltatni a Gemini adathalmazokat.

2.3. Egységes munka a különböző emberi tapasztalatokkal

A munka jellegéből adódóan a chatbotok válaszainak értelmezése az azt olvasó, értelmező személytől nagyban függ. Az emberi tapasztalatok, szakterületi tudások eltérhetnek egymástól; ennek kiküszöbölésére munkánk során több ember segítségét is igénybe vettük. A generálás folyamatát, illetve a kimeneti szövegek jellemzését több ember egységesített véleményéből, észrevételéből kaptuk meg. A Gemini adathalmazon a dolgozat egyik szerzője, illetve egy nyelvész végzettségű személy dolgozott, a ChatGPT-vel való munkát a másik szerző, egy nyelvészszakértő, egy informatikus kutató, valamint egy mesterszakos hallgató végezte. A közös munka segítségével hatékonyabban tudtuk

finomhangolni a promptok vázát, valamint sokkal több észrevételt voltunk képesek dokumentálni: ezen eredmények rendre a 3.1, valamint 3.2 és 3.3 fejezetekben kifejtésre kerülnek.
Habár a szövegek generálásában és az emberi tapasztalatok leszűrésében több munkatársunk is tevékenyen részt vállalt, a kísérletek felépítését, a modellek és nagy részben a szükséges szkriptek implementálását, a kísérleti eredmények mérését, és következtetéseik leszűrését a dolgozat két szerzője végezte.

3. Saját kísérletek ember, ChatGPT, Gemini szövegek elkülönítésére

3.1. Adat, jellemzőkészlet bemutatása

A kísérletek során kezdetben nem állt rendelkezésünkre Gemini által generált adat, ezért csak emberi, illetve ChatGPT által generált szövegekkel dolgoztunk. Az adatunk ebben a kezdeti szakaszban 500 emberi szövegből állt és 500 ChatGPT 4.0 által generáltból. Az emberi és gépi szövegek is "különböző szerzőktől" származtak, annak érdekében, hogy több szövegstílusból is tanulhasson a modell, és így általánosságban legyen képes elválasztani a két osztályt, ne csupán 1-2 stílusra tanuljon rá. A gépi szövegek esetén a "különböző szerzős" szövegek generálása érdekében generált személyiségeket alkalmaztunk. Minden prompt esetén egy másik személyiséget adtunk a gépnek, különböző személyiségjegyekkel. Alap adataink elkészítésénél a promptunk váza az alábbi volt, GPT és Gemini esetén is:
Viselkedj úgy, mintha ez az ember lennél, akit alább részletezek:
SZEMÉLYISÉG LEÍRÁS KEZDÉS
PERSONA
SZEMÉLYISÉG LEÍRÁS VÉGE
Ezzel a jellemmel írj 1db legalább 600 szavas magyar fórum kommentet. Csak és kizárólag maga a szöveg érdekel, ne emeld ki külön, hogy mi a téma, vagy ki írta a szöveget, vagy bármi egyebet. Legyen minimum 600 szavas. Ne írd bele a szövegbe a személyes adataidat, amiket fentebb részleteztem, például ne

írd bele a neved vagy a lakcímed.
A téma lehet szerelem, étel, ital, turizmus, politika, játékok, filmek, sorozatok, zenék, függőségek, traumák, szexualitás, öltözködés, oktatás, egészségügy, filozófia, férfiak, nők, álláskeresés, programozás, mesterséges intelligencia, bűncselekmények, természeti katasztrófák, emberek, állatok, növények, gyerekek és házasság is.
The text must not contain any ordered lists or unordered lists. A szövegben ne legyen lista, se rendezett se rendezetlen.
A szöveg legyen véletlenszerű, olyat generálj amit korábban még nem generáltál. Ne legyen címe a kommentnek. Több mint egy soros legyen az első bekezdés.
Ezt a promptot iteratív módon állítottuk elő a nagyobb hibák figyelembe vételével, arra törekedve, hogy emberi szemmel meggyőző kimenetet kapjunk mindkét modellel. A szövegek hossza a prompt ellenére is inkonzisztens volt és tapasztalatok szerint se a ChatGPT, se a Gemini nem tartotta be az alsó szóhatárt, amit a prompt kért tőle. Majdnem minden esetben szignifikánsan rövidebb szöveget adott mindkét modell a kértnél. A prompt így bár 600 szót kért a modelltől, átlag 300-400 token között volt a szöveg hossza. Ez a működés potenciálisan egy beleégetett szabály: a ChatGPT annak API-jának használata során szódarabok (wordpiece) alapján számláz, ezért lehetséges, hogy mindkét modell a szódarab menti mennyiséget veszi elsődlegesen figyelembe - bizonyos értelemben a szódarabok szolgálhatnak tokenként, amik azonban rövidebbek, mint a tényleges tokenek. Történtek olyan próbálkozások is, ahol definiáltattuk a modellel, hogy mit jelent a token kifejezés (vagy a szó), illetve olyan is volt ahol külön leírtuk a promptban azt is hogy ez pontosan mit jelent, viszont minden próbálkozás sikertelennek bizonyult.
A témákat tapasztalat alapján szükséges volt felsorolni, konkrét példákat adni, ugyanis, ha "általános szöveget" kértünk a téma meghatározása nélkül, a szövegek még rövidebbek voltak, és kevésbé változatosak is. Szükség volt arra is, hogy a listákat angolul és magyarul is kizárjuk a promptban tapasztalatok szerint, máskülönben gyakran tett formázva listákat a szövegbe, amik embertől szokatlanul formálisak lennének. Így is előfordult azonban, hogy került lista a szövegbe, ezeket teljes mértékben nem tudtuk kiszűrni a prompttal.
A fent leírt sémában a "PERSONA" helyére került az éppen használt szemé-

lyiség leírása, egy személyiség példa:
** Név: A. Balogh Ádám*
** Lakcím: Ady endre utca 214.*
H-7611 Egyházasmártonfa
** Munka: Fizikus at Oláh Bt.*
** Születési dátum: 1978-12-22*
** Végzettség: 8 általános*
** Személyiség jegyek:*
- bíráló
- tapintatos
- elõrelátó
- boldog
- rideg
- egocentrikus
- empatikus
- lelkesítõ
- ambiciózus
- intelligens

A személyiségeket véletlenszerűen generáltuk egy munkatárs segítségével. A személyiségjegyeket különböző kategóriákba osztottuk, ezek: jó személyiségjegyek, pozitív személyiségjegyek, vezetői személyiségjegyek, negatív személyiségjegyek, semleges személyiségjegyek.

Minden személyiség kapott véletlenszerűen 3 jó személyiségjegyet, 2 pozitív személyiségjegyet, 1 vezetői személyiségjegyet, 3 negatív személyiségjegyet, illetve 2 semleges személyiségjegyet. Ezen felül minden 4. személyiség rendelkezett a rasszizmus személyiségjeggyel is. Az eloszlást ezen személyiségjegy esetén rögzítettük, azon okból hogy előfordulása másképpen túl ritka lett volna a többi személyiségjegyhez képest.

Az alábbiak tartoztak a jó személyiségjegyek közé: kiegyensúlyozott, harmonikus, nyugodt, békés, boldog, vidám, művelt, tanult, széles látókörű, okos, intelligens, eszes, kreatív, találékony, megfontolt, előrelátó, tervező, önálló.

A pozitív személyiségjegyek: közvetlen, barátságos, megbízható, lelkiismeretes, őszinte, igaz, udvarias, tapintatos, előzékeny, megértő, empatikus, elkötele-

zett, hűséges, gondoskodó, támogató, gyakorlatias, pontos, precíz, aprólékos, szorgalmas, kezdeményező, vállalkozó, humoros.
Vezetői személyiségjegyek: felelősségvállaló, motiváló, inspiráló, lelkesítő, irányító képességű.
Negatív személyiségjegyek: agresszív, erőszakos, durva, kritikus, bíráló, lekezelő, szórakozott, figyelmetlen, szétszórt, igénytelen, szeszélyes, kiszámíthatatlan, számító, érzéketlen, rideg, zsarnoki, despota, zsugori, fukar, élősködő, lusta, tétlen, önző, egocentrikus, rosszindulatú, gyáva, képmutató, álszent, köpönyegforgató.
Semleges személyiségjegyek: meggyőző, szenvedélyes, ambiciózus.
Ezen felül minden személyiség kapott egy végzettséget is az alábbi 5-ből: 8 általános, érettségi, szakmunkás, főiskola, egyetem.
Az adatok teljes mértékben fiktívek, beleértve a nevet, a lakcímet, a munkahelyet és a születési dátumot is.
A kísérletek során az ember által írt szövegeket a Hunglish 1.0 korpuszból [35] nyertük ki. A felhasznált emberi szövegek magyar nyelvű, hétköznapi nyelvezetű fórum kommentek az index.hu elődjéről. A ChatGPT 4.0 által generált szövegeket annak API-ja segítségével generáltuk, míg a Gemini által generált szövegeket kézzel kényszerültünk legeneráltatni. Ahogy korábban az említve volt, a Gemini API-ja Magyarországon nem elérhető. Egy munkatársunk az online Gemini felületen minden promptunkat sorban bemásolta a chat ablakba, minden prompt után új beszélgetést indítva. Tapasztalataink szerint, ha egy beszélgetésen belül próbáltunk több üzenetet is generáltatni, az üzenetek változatossága drasztikusan romlott, hajlamos volt a modell nagyon hasonló szöveget generálni az első generált szövegéhez minden alkalommal.
A későbbiekben, rendelkezésünkre állt 486db, Gemini által generált szöveg is, melyeket ugyanazon promptokkal generáltunk, mint a GPT-s szövegeket is. Ezt kiegészítettük további 499 GPT által generált szöveggel, illetve 500 emberi szöveggel. Végül a korpuszunk 1000db emberi, 999db ChatGPT 4.0-ás, illetve 486db Gemini által generált szövegből állt. Az adathalmazok szövegeinek hosszát az 1. táblázatban részletezzük.

1. táblázat. Karakterek és tokenek száma a végső adathalmazban

	Emberi	**ChatGPT**	**Gemini**
Átlagos karakterszám	1844	2550	2215
Átlagos tokenszám	295	367	307
Minimum karakterszám	386	1053	1272
Maximum karakterszám	6684	4384	4365
Minimum tokenszám	49	154	165
Maximum tokenszám	1042	679	624
Karakterszám szórása	593	447	460
Tokenszám szórása	91	61	61

Belátható, hogy a szórások fedik egymást mindhárom osztály esetén, mind tokenszámot, mind karakterszámot illetően. Átlagban a ChatGPT által generált szövegek voltak a leghosszabbak, illetve megállapítható, hogy a gépi szövegek szórása jelentősen alacsonyabb volt az emberinél mind tokenszám, mind karakterszám tekintetében. Ennek ellenére is magasnak bizonyult azonban, azaz elmondható, hogy nem követte egyik modell sem konzisztensen a promptban kért szöveghosszt.

Kísérleteink során törekedtünk minél emberibb szövegeket generálni, így a később generált 499db GPT szöveg prompt váza eltért a korábbiaktól, azonban továbbra is alkalmazva voltak személyiségek. A prompt váza az alábbi volt minden esetben:

A következő személy nevében írj egy fórum hozzászólást:

PÉLDA: "Daddy, azt hiszem, itt a kurva vég! Az árak egyenesen az égbe szöknek az élelmiszerek terén! Nem lehet normális áron hozzájutni semmihez! Az istenit, még a kenyér ára is a csillagokban jár! És a tej? Kurva drága, mint a gecibe! Nem tudom, hogy képzelik ezt az emberek, de már tényleg sok! De nem hagyom magam, Daddy! Megvan a tervünk: felszabadulunk a boltok rablástól, és irány a hegyekbe! Felveszünk pár piát, és eltűnünk a kurva világból egy kis időre! Túl sokat basztak már meg minket, ideje, hogy mi is visszaüthessünk! Holnap összepakolom a cuccot, aztán irány a vadonba! De előtte még fszozzunk ki egy kicsit az áruházban, mert megérdemeljük a jó kis lökést! Tudom, hogy te is benne vagy, mert te vagy az a kurva legenda, aki mindig az oldalamon áll! Fszom, úgy érzem, hogy ez lesz az, amire mindketten vártunk! Kurva jó

lesz, nemde? Tudod mit, még egy kis döglesztést is beiktatunk közben! Igenis, mesterek vagyunk mi!"
SZEMÉLY: IDE JÖN A PERSONA
TÉMA: IDE JÖN A TÉMA
HANGVÉTEL: IDE JÖN A HANGVÉTEL
*HYPER PARAMÉTEREK: * ne említsd a nevét és végzettségét * helyezz el j-ly elírásokat * language: hungarian * no pragraph separation * atleas 600 word length * frequent use of swear words * do not quote from the example * forum comment * no paragraph separation * nincs bekezdésekre bontás * diverge from the topic of the example * address only once * only one paragraph, which is atleast 25 lines long.*
A személy alá itt is minden esetben egy random generált személyiség került ahogy az fentebb is részletezve volt. A téma az alábbiak közül volt véletlenszerűen kiválasztva: szerelem, étel, ital, turizmus, politika, játékok, filmek, sorozatok, zenék, függőségek, traumák, szexualitás, öltözködés, oktatás, egészségügy, filozófia, férfiak, nők, álláskeresés, programozás, mesterséges intelligencia, bűncselekmények, természeti katasztrófák, emberek, állatok, növények, gyerekek, házasság.
A hangvétel szintén véletlenszerűen volt kiválasztva, az alábbiak közül: vidám, szomorú, haragos, unott, izgatott, lenéző, félénk, szerény, tapintatos, durva, fennkölt, fenyegető, hivatalos, nyugodt.

3.2. Észrevételek a Gemini működése kapcsán

A Gemini esetében, mivel kézzel történt az adatok generálása, ezért sok működési elvet a korábban említett nyelvész kollégánk segítségével figyeltünk meg; ezek feloldása, kikerülése további kihívásokkal jelentett. A modell nagyon sokat írt a mesterséges intelligencia és a boldogság témákról, a szövegek megközelítőleg 65%-ánál a két téma valamelyikét használta fel, mint fő tartalom. Általában pozitív hangvétellel, logikai szempontból enyhe naivitással rendelkező szövegeket generált. Nagyon sok esetben ugyanazt a sémát használta kis módosítások mellett, például: kinéz az ablakon - kávé gőzölög - elgondolkodik egy adott témán / elgondolkodik a jövőn - technológia szerepén - AI / régen

minden jobb volt. Negatív hangvételű szövegek nagyon ritkán generálódtak. Ha a prompt tartalmazta a "rasszista" szót (mint személy jellemzőjét), akkor hibaüzenetként jelezte, hogy rasszista szövegeket nem generálhat. Bekezdéseket jellemzően nem emberszerűen tagolta a modell. Sok esetben előfordult, hogy a megadott prompt valamelyik részét figyelmen kívül hagyta: személyes adatok említései benne maradtak, egyéb utasításoknak nem tett eleget, stb. Bizonyos személyiségek esetén a háló képtelennek bizonyult megfelelően szöveget generálni: az adott promptra adott válasz többszöri újra generálása után is a Gemini "nem tudott" a megadott utasítások szerint eljárni semmilyen módon. Több promptban is kifejezetten kértük a hálót elgépelésekre, rossz ragozásra, szavak fel- és kicserélésére, azonban a helyesírási és mondattani hibák teljes mértékben hiányoznak a szövegekből. Nem létező szavakat látszólag képtelen generálni. Egyedüli hiba, melyet sokkal konzisztensebben képesek voltunk generáltatni, az a mondat, gondolat akár szó közepén történő be nem fejezése (például: "[...] végülis a magu"), azonban ez sokkal inkább köthető a chatbot működéséhez. Előfordult, hogy egy választ elkezdett megírni a modell, majd azt kitörölte, és egy merőben más végső választ alkotott meg. Arra is volt példa, hogy egyszerűen csak abbahagyta a szöveggenerálást, kiszámíthatatlan pontokon. Utóbbi viselkedés a ChatGPT esetében is megtapasztalható. Ezek mind működési sajátosságok miatt alakultak így, nem pedig azért, mert a promptban megadott utasítások mentén járt volna el. Logikailag nem mindig kapcsolódnak egymáshoz a témák, például: AI szerepe napjainkban a téma, amit egyből egy recept diskurzusa követ. Ez a működési elv nem feltétlen tekinthető hibásnak, az ADHD-val diagnosztizáltaknál például gyakran előforduló jelenség a gondolatinkoherencia. Nagy leterheltség, hosszabb ideig – nagyjából másfél óráig – tartó szöveggenerálás esetén a modell elkezdi a téma szempontjából ismételni magát; további terhelés után pedig különböző hibaüzeneteket ír, melyekkel közli a felhasználóval, hogy nem áll módjában több/ilyen tartalmat generálni, vagy teljesen "lefagy" ("Gemini is on a break. Come back in a second."), leáll a működése. Várakozás után a problémák jellemzően enyhülnek. Ezen felül nyelvész kollégánk további, szociolingvisztikai [22] szempontból releváns jelenségre hívta fel a figyelmet: a magasabb társadalmi megbecsültséggel rendelkező foglalkozások (többek között orvos,

ügyvéd, egyetemi tanár) szinte mindig a mesterséges intelligenciáról beszélnek. Ennek ellentétje, például megszemélyesített szakmunkások szintén foglalkoznak a témával, viszont általában indulatosabb hangvételt alkalmazva – az indulatos megnyilvánulási forma általánosságban is megjelenik a Gemini által generált megszemélyesítésekben. A helyesírási hibák mentességéhez hasonlóan a modell képtelen olyan módon és annyira helyesen, komplexen írni, fogalmazni, mint ahogy a prompt alapján megtestesítendő személyre az igaz lenne; a mesterséges intelligencia nem tesz különbséget a korcsoport, foglalkozás, végzettség, nem, lakhely és személyiségvonás körében, minden személyhez ugyanazt a fogalmazási készséget rendeli hozzá. Bizonyos promptok feldolgozása esetén, ha a modell nem tudta "értelmezni" a bementet, akkor ezt jelezte, és figyelmünkbe ajánlotta a Google keresőjét, mint potenciális platform az információ gyűjtésre; nagy valószínűleg ez is egy szabályalapú működési elv a háló esetében.

3.3. Észrevételek a ChatGPT működése kapcsán

A ChatGPT esetében is történt kézi szöveggenerálás, ezek a szövegek képezik a korábban említett 28, illetve 60 szöveges teszthalmazainkat. A 60 szöveget a nyelvvészszakértő segítségével állítottuk össze. A generálás folyamán a nyelvvészszakértőnek nem mutattunk minta promptot, csupán azt adtuk meg célnak, hogy legalább 600 szóból álljanak a szövegek, illetve mimél emberibbek legyenek. A folyamat során a ChatGPT működésével kapcsolatban is megfigyeltünk bizonyos működési mintázatokat, szabályosságokat - ezek egy része hasonló módon megfigyelésre került a Gemini esetében is: a politikailag korrekt szöveg szűrése látszólag nem teljesen determinisztikus, előfordul, hogy "rasszista" jellemmel is generál szöveget a modell, ám ez az esetek kisebb hányadában fordul csak elő. Felmerülhet a kérdés, hogy a modellek végsőnek generálandó szövegei egy bemeneti szűréstől függenek-e, ami feltehetőleg nem determinisztikusan történik, vagy a kimeneteket determinisztikusan szűrik, ami azonban nem mindig megfelelő. Nagyon nehéz rávenni a neurális hálót, hogy ne általános, filozofikus szöveget írjon. Kitalált személy és helyszín nevek kért említése esetén is csak ritkán jelennek meg ilyen nevek a szövegben. A

promptokban szereplő utasítások közül némelyeket a modell spontán figyelmen kívül hagy. A Geminihez hasonlóan szintén nehezen tagol emberszerűen bekezdésekre. Fórumhozzászólás írása esetén szinte minden esetben 3-6 sornyi szöveget állít elő bekezdésenként. ChatGPT esetén egyes esetekben sikerült bekezdésmentes szövegeket generálni, ám azok tartalmilag nem voltak megfelelőek. A sor- és bekezdésszám nem tartásában persze szerepet játszhatnak külső tényezők is, mint például a felhasználó képernyőjének mérete, azonban az elvárt szó- és tokenmennyiség esetén is hasonló hibákat tapasztaltunk. Ha megadtunk egy konkrét jellemet, akkor ennek jellemzőit (például kedves, fölényes, rasszista) alapból megpróbálta explicit módon belefogalmazni a szövegbe, ha pedig megkérjük, hogy ne fogalmazza bele a jellemet, csak használja fel azt (például elmondjuk, a személy nem tud róla, hogy ezek a jellemzői, a jellemvonások csak és kizárólag a szöveg stílusában látszanak meg), akkor nem kerülnek említésre a jellemvonások, ám a szövegből látszólag hiányzik az implicit beleillesztése is a jellemzőknek. Látszólag a ChatGPT is képtelen helytelenül, nem létező szavakat írni. Ha a promptban kifejtjük, hogy a szöveget író személy nagyon rosszul fogalmaz, és nagyon rossz minőségű írást produkál, akkor a tapasztalatok szerint köznyelvibb (ezzel egyúttal pedig sokszor élethűbb) megfogalmazást tudunk elérni a modellel. Ennek ellenére a Geminihez hasonló módon itt sem tapasztaltunk kitalált, vagy akár csak elírt szavakat, függetlenül a megtestesítendő személyiségtől.

3.4. Kísérletek felépítése

A kísérletek során minden esetben 10-szeres ("10-fold") keresztvalidációt [19] alkalmaztunk kiértékelésre. A tanító halmazba minden esetben az adat 90%-a került, míg a maradék 10% alkotta a validációs halmazt. Minden konfiguráció esetén 10 kísérlet történt, olyan módon, hogy a 10 kísérlet során minden alkalommal másik 10% volt a validációs halmaz (a maradék 90% pedig a tanító halmaz), melyek között átfedés nem volt. Ilyen módon minden konfiguráció kiértékelésre került a teljes adathalmazon. A konfigurációkat első sorban a 10 foldon elért átlag Macro F1 pontszám alapján hasonlítottuk össze, de a cikkben ki fogunk térni a legjobb modellek esetében az 1-1 foldon elért

eredményekre is. Minden konfiguráció esetén pontosan ugyanolyan módon voltak elosztva a foldok, így két konfiguráció eredménye azonos foldon minden esetben összehasonlítható.
A validációs adathalmaz (a tanító nem) bizonyos kísérletek esetén utófeldolgozásra került két különböző módszer közül az egyikkel. Az egyik esetben szövegbe véletlenszerűen elírásokat szúrtunk, olyan módon, hogy adott véletlenszerűen választott karaktereket arra a karakterre cseréltük, amely egy magyar nyelvű billentyűzeten az adott karakter mellett szerepel. Például "w" helyett "q" vagy "e", illetve "q" helyett "w". Ha egy billentyű mellett két másik billentyű is szerepelt, mint például a "w" esetén, abban az esetben véletlenszerűen döntöttünk arról, melyik szomszédjára cseréljük le az eredeti karaktert. Célunk az ilyen stílusú elírások beszúrásával az emberi félregépelések szimulálása volt, ezzel tesztelve, hogy a modellünk (ami az eredeti, nem utófeldolgozott tanítóhalmazon tanult) milyen mértékben támaszkodik helyesírási jellemzőkre, és milyen szinten rontaná a predikció pontosságát az, ha valaki a gépi generált szövegbe szándékosan elírásokat tenne annak érdekében, hogy azt higgyék mások, hogy ember írta. Az elírások száma szintén véletlenszerű volt minden szöveg esetén, megközelítőleg minimum 90, maximum 110 volt. Bizonyos esetekben az elírások száma kis mértékben kívül eshetett ezen az intervallumon, azonban nem szignifikáns mértékben (például, ha már eleve volt benne elírás, bár gépi szövegeknél ez kifejezetten ritkán fordult elő). Kezdetben csak kis mennyiségben szúrtunk be elírásokat (például körülbelül 10-et), így bár feltehetően közelebb volt az elírások száma ahhoz, amit az átlagember hasonlóan hosszú gépelt szövegében találnánk, a teljesítmény nem változott szignifikánsan. Végeztünk kísérleteket olyan módon is, hogy csak a GPT, illetve a Gemini által generált szövegekbe szúrtunk elírásokat, illetve úgy is, hogy az emberi szövegekbe is.
Az említett adathalmazok mellett a legnagyobb pontosságú modelljeinket kiértékeltük független teszthalmazokon is. Az alábbi teszthalmazokon értékeltünk ki:

- 28db ChatGPT 4.0 által generált szöveg
- 63db emberi szöveg a ParlaMint 4.0 [6] korpuszból

- 60 db vegyes, ChatGPT 3.0 és ChatGPT 4.0 által generált szöveg

A 28db ChatGPT 4.0 által generált szöveg esetében hasonló promptokat alkalmaztunk, mint a korábbi adathalmazoknál, kisebb módosításokkal: mind a tanító-, mind a validációs halmaztól eltérő stílusú szövegeket generáltunk annak érdekében, hogy általánosabb képet kapjunk modelljeink teljesítményéről. Ezek a promptok a mellékletben 6 olvashatók. A ParlaMint korpusz több szövegből is áll mint 63, azonban nyelvi elemzőnkkel több szöveg elemzésére a kutatás során nem jutott már kapacitás, így a szövegek maradéka esetén csak a szövegbeágyazásokat használhattuk volna fel. A 60db vegyes, GPT-s adathalmazunk szövegei nyelvész által voltak generáltatva a dolgozatban taglaltaktól nagy mértékben eltérő promptokkal, kifejezetten azzal a céllal, hogy emberinek tűnjenek. Különböző típusú szövegeket tartalmaz, köztük ál-bombafenyegetéseket, kitalált párbeszédeket, illetve kitalált meséket. Stílusban nagyban eltér tehát a többi szöveghalmaztól.
A 28 GPT-s szövegben a szöveghosszakra az alábbiak teljesültek:

- az átlagos karakterhossz 4353, szórása 400
- az átlagos tokenszám 608, szórása 48

Míg a 63 ParlaMint szövegekre az alábbiak:

- az átlagos karakterhossz 434385, szórása 181933
- az átlagos tokenszám 57496, szórása 24259

A 60 további GPT-s szövegekre az alábbiak:

- az átlagos karakterhossz 1348, szórása 893
- az átlagos tokenszám 195, szórása 129

A ParlaMint szövegek tehát jelentősen nagyobb terjedelműek, mint a többi korpusz szövegei. A kiértékelés során ezt figyelembe is vettük, erről később a "Kísérleti eredményekben" szekcióban lesz részletesebben szó. A 60 ChatGPT szöveges teszthalmazban pedig rendszerint rövidebb szövegek voltak.

A használt jellemzők a HuBERT szövegbeágyazásai voltak, illetve a nyelvi elemzőnkből kinyert jellemzők. Az utóbbiakat több kategóriába is soroltuk, ezek a dolgozat végén a "Nyelvi jellemzők" szekcióban találhatóak részletesebben kifejtve.

3.5. Adatok elemzése

Adathalmazaink elemzése során megállapítottunk jelentős különbségeket bizonyos jellemzők előfordulásai arányaiban: például egy jelentősebb eltérés a névelemek előfordulásai arányai voltak. Az emberi szövegekben drasztikusan gyakrabban fordultak elő névelemek, mint gépi szövegekben.
Mindegyik jellemző esetén az emberi teszthalmazhoz viszonyítottuk két modell jellemzőnkénti eltéréseit. A jellemzőarányokat három lépcsős szűrés keretén belül vizsgáltuk: az első lépcső esetében pusztán a relatív és abszolút eltérések kerültek feljegyzésre, valamint azt is jeleztük, ha az adott jellemző nem szerepel emberi szövegekben (a táblázatban no_index cellaérték felel a jelzésért), Gemini szövegekben (no_gemini), ChatGPT szövegekben (no_GPT), illetve, ha egyik esetben sem íródott az adott jellemzőből (no_index_and_gemini és no_index_and_gp). A relatív eltérés esetén az értékek azt jelzik, hogy hány százalékkal kisebb az adott jellemzőarány a megfelelő modell által generált szövegekben az emberiekhez képest (tehát a GPT oszlopban a -15-ös érték azt jelenti, hogy a ChatGPT 15%-kal nagyobb arányban generálta az adott jellemzőt). A második lépcső esetén a 10%-nál nagyobb eltéréseket vettük csak figyelembe. A harmadik szűrési szinten egyedül a csupa nemnulla értékkel rendelkező sorokkal dolgoztunk. A szűrt adattáblák megtalálhatóak a mellékletben 6 rendre big_result_big.csv, big_result_medium.csv és big_result_small.csv néven – a result_ kezdetű adattáblák a névelemenkénti eltéréseket tartalmazzák, ám ott mind a három adattábla megegyezik.

3.6. Jellemző-kombinálási módszerek és multimodalitás

A 2. táblázat a jellemző-kombinálási csomag [9] azon kombinálási módjait mutatja be, melyeket kutatómunkánk során mi is felhasználtunk. A jellemző-kombinálási módok kérhetnek numerikus és/vagy kategorikus jellemzőket, az

általunk használt felhasználási módok esetén mi csak numerikus adatokat adtunk meg, a jellemzőkészletünkben nem voltak jelen kategorikus jellemzők. Numerikus jellemzőknek számítunk valós szám értékeket, például 0.1, 0.2, 0.3, stb, míg kategorikus jellemzőnek számítunk diszkrét értékkészletű értékeket, például lehet 1-es kategória, 2-es kategória és 3-as kategória, az értékek pedig 1, 2 illetve 3.

2. táblázat. Jellemző-kombinálási módszerek összegzése

Jellemző-kombinálási mód neve **Leírás**
text_only Csak a HuggingFace transzformer által feldolgozott szövegoszlopokat használja a végső osztályozó réteg(ek) előtt. Lényegében megegyezik a HuggingFace ForSequenceClassification modelljeivel
concat Együttesen összevonja a transzformer kimenetét, a numerikus és kategorikus jellemzőket a végső osztályozó réteg(ek) előtt
individual_mlps_on_cat_and_numerical_feats_then_concat Külön MLP-t alkalmaz a kategorikus és numerikus jellemzőkre, majd azokat együttesen összevonja a transzformer kimenetével a végső osztályozó réteg(ek) előtt
attention_on_cat_and_numerical_feats Figyelem (attention) alapon kérdezi le a transzformer kimeneteket és a numerikus, valamint kategorikus jellemzőket a végső osztályozó réteg(ek) előtt
gating_on_cat_and_num_feats_then_sum Kapuzott összegzést alkalmaz a transzformer kimenetekre, a numerikus, illetve a kategorikus jellemzőkre a végső osztályozó réteg(ek) előtt
weighted_feature_sum_on_transformer_cat_and_numerical_feats A transzformer kimenetei, valamint a kategorikus és numerikus jellemzőket a tanult súlyok alapján összegzi a végső osztályozó réteg(ek) előtt

A jellemző-kombinálási csomag szövegek kategorikus és numerikus jellemzőinek különböző kombinációjú felhasználását teszi lehetővé. Segítségével az adatnak több különböző változatán voltunk képesek tanítási kísérleteket végezni. A kombinált multimodális jellemzőket **m** jelöli, **x** a transzformer kimeneti szövegjellemzőit jelöli, **c** jelöli a kategorikus jellemzőket, **n** pedig a numerikus jellemzők jelöléséért felel. **W** felel a súlymátrixért, **b** pedig a skaláris torzítást jelöli.

Az **MLP** (Multilayer Perceptron - Többrétegű Perceptron) modell egy olyan előrecsatolt neurális háló modell, mely teljesen összekapcsolt, a kapcsolódások között pedig nemlineáris aktivációs függvények szerepelnek. Struktúrájából adódóan a modell képes nemlineáris kapcsolatban álló adatok szétválasztására, feldolgozására. A kísérletekhez felhasználandó jellemzőkre teljesül ez a típusú nemlinearitás, így az MLP modell alkalmazása ez esetben kézenfekvő.
A jellemzőket, valamint a többi adatot mielőtt átadtuk volna a modul egyik kombinálási módszerének, először előfeldolgozás gyanánt normalizáltuk. A kombinálási módszerek során maga az összevonás, illetve az adott esetben felhasznált jellemzők halmaza is változott. Csak szöveg (jellemzők nélküli) felhasználása – **text_only** – esetén:

$$m = x$$

jól látható, hogy csak és kizárólag szöveg alkalmazása esetén a multimodális jellemzők megegyeznek a transzformer kimeneti szövegjellemzőivel; sem numerikus, sem pedig kategorikus jellemző nem került felhasználásra, így nem történik változás a szövegjellemzők halmazán.
Az egyszerű összefűzés (**concat**) módszer használata esetén nem történik – a normalizálást leszámítva – előfeldolgozás egyik kategóriát illetően sem, egyszerűen a szöveg végére fűzzük a jellemzőket a szöveg beágyazása előtt.

$$m = x \parallel c \parallel n$$

A következő módszer (individual MLPs on categorical and numerical features then concat; röviden MLP + concat) esetén előfeldolgozzuk a numerikus jellemzőket, melyeket a szintén előfeldolgozott kategorikus jellemzőkkel, valamint a transzformer kimeneti szövegjellemzőivel vonunk össze (**individual_mlps_on_cat_and_numerical_feats_then_concat**).

$$m = x \parallel MLP(c) \parallel MLP(n)$$

A jellemzőknek, valamint a neurális hálók működésének köszönhetően másmás eredményt kapunk, ha az adatokat több kis részben (batch), vagy egyben

dolgozzuk fel. Ezen működés kiküszöbölésére szolgál az a módszer, melynek segítségével egyben előfeldolgozzuk a numerikus és kategorikus jellemzőket, a kapott eredményt pedig összefűzzük a transzformer kimeneti szövegjellemzőivel (**mlp_on_concatenated_cat_and_numerical_feats_then_concat**).

$$m = x \parallel MLP(c \parallel n)$$

Előfordulhat, hogy pusztán az adatok előfeldolgozása, majd összefűzése nem eléggé pontos, torz eredményekhez vezetve. Szükséges lehet a jellemzők valamilyen metrika mentén történő súlyozása, korrigálása. Ezt teszi lehetővé a következő módszer (**attention_on_cat_and_numerical_feats**), mely a jellemzők ön-figyelmét használja fel. Az ön-figyelem (self-attention) modul a bemeneti vektorból egy lekérdezés (query), kulcs (key) és érték (value) vektort állít elő. Ez után a modul egy hasonlósági metrikát alkalmazva meghatározza a hasonlósági mértéket a lekérdezés és kulcs vektorok között. Ezzel a mértékkel módosítjuk az érték vektor elemeit, majd ezen módosított vektor elemeinek összegzése lesz a modul kimenete.

$$m = \alpha_{x,x} W_x x + \alpha_{x,c} W_c c + \alpha_{x,n} W_n n$$

A W súlymátrixok első dimenziója a transzformer kimeneti szövegjellemzőinek száma, második dimenziója pedig a megfelelő jellemző számossága; például a W_n a (transzformer kimeneti szövegjellemzőinek száma, numerikus jellemzők száma) dimenziójú súlymátrixot jelöli. A figyelem mátrix $\alpha_{i,j}$ együtthatója a következő módon határozható meg:

$$\alpha_{i,j} = \frac{exp(LeakyReLU(a^T[W_i x_i \parallel W_j x_j]))}{\Sigma_{k \in \{x,c,n\}} exp(LeakyReLU(a^T[W_i x_i \parallel W_k x_k]))}.$$

Először egy vektorrá fűzzük össze a megfelelő transzformer kimeneti szövegjellemzők és súlyvektorok kompozícióit – a számlálóhoz az aktuális i. és j. vektort használjuk fel, a nevezőben az aktuális i. vektor, valamint az összes jellemzővektor felhasználásra kerül. Az **a** vektort, ami a jellemzők figyelmének vektora, összeszorozzuk a kapott vektorral, az eredményt pedig a **LeakyReLU** függvénnyel feldolgozzuk. A **ReLU** (Rectified Linear Units -

Helyesbített Lineáris Egységek) egy olyan aktivációs függvény, mely a pozitív tartományban pozitív értéket, a negatív tartományban pedig nulla értéket vesz fel. A függvény érdekessége a nemlinearitás. A pozitív tartománybeli linearitás azzal a vonzó tulajdonsággal rendelkezik, hogy megakadályozza a gradiens telítetlenségét (ellentétben a szigmoid aktivációs függvényekkel), ám a valós számegyenes felén a gradiens nulla értéket vesz fel. A ReLU aktivációs függvény megvalósítása:

$$f(x) = max(0, x).$$

A LeakyReLU [27] a negatív tartomány kezelésében tér el az egyszerű ReLU függvénytől. Negatív értékek esetén a konstans nulla félegyenes helyett egy lejtővel dolgozik, melynek együtthatója a tanulás előtt előre meghatározott. A függvényt ritka gradienssű feladatoknál érdemes használni.

$$f(x) = \begin{cases} x, & \text{if } x >= 0 \\ ax, & \text{egyébként} \end{cases}$$

A **gating_on_cat_and_num_feats_then_sum** nevű módszer segítségével lehetőségünk van kapuzott jellemzők összegzését felhasználni. Ezt a módszert az Integrating Multimodal Information in Large Pretrained Transformers [28] cikkben használt kapuzási módszer inspirálta – érdemes lehet megjegyezni, hogy ugyanez a cikk szolgál a jellemző-kombinálási csomag egyik alapjául is.

$$m = x + \alpha h$$

$$h = g_c \odot (W_c c) + g_n \odot (W_n n) + b_h$$

$$\alpha = min(\frac{\parallel x \parallel_2}{\parallel h \parallel_2} * \beta, 1)$$

ahol β egy hiperparaméter, W_c és W_n továbbra is a transzformer kimeneti szövegjellemzőinek, valamint a megfelelő típusú jellemzők számosságából

alkotott súlymátrixok. A kapuzási vektort g jelöli, g_i értékét pedig egy R aktivációs függvény határozza meg,

$$g_i = R(W_{g,i}[i \parallel x] + b_i)$$

ahol W_{gi} egy (transzformer kimeneti szövegjellemzőinek száma, i. jellemző számossága + szöveges jellemzők számossága) dimenziójú súlymátrix.

4. Kísérleti eredmények

4.1. Alap jellemzőhalmaz, kezdeti eredmények

Minden kísérlet esetén – kivéve ahol csak szövegbeágyazásokat használtunk – használtuk az olvashatósági metrikákhoz kapcsolódó jellemzőinket, ezekre később "alap jellemzőhalmazként" fog hivatkozni a dolgozat. Egyéb jellemzőket ezek mellett használtunk.
A kezdeti kísérletek során, amikor még csak 500 GPT-s, illetve 500 emberi szöveg állt rendelkezésünkre, ezért az osztályozási feladat csak két osztállyal történt, a feladat láthatóan triviálisnak bizonyult a modellnek, több konfiguráció esetén is tökéletes (100%-os pontosság mind a 10 validációs foldon) eredményeket értünk el, például többek között ha az alábbi jellemzőket használtuk: *fleschReadingEaseScore, gunningFogIndex, fleschKincaidGradeLevel, colemanLiauIndex, smogIndex, automatedReadabilityIndex, linsearWriteIndex, wordBeforeHesitationRate, repeatRate, restartRate, smileyRate, quoteRate, wordAfterHesitationRate, dashRate, dateRate, hesitationRate.*
Azaz az alap jellemzőhalmazunkat, illetve a pragmatikai jellemzőinket használtuk (lásd: Nyelvi jellemzők szekció). Fontos megjegyezni, hogy a jellemzők közül több jellemző értéke is lehetett 0 a szövegek többségénél, így feltételezhetően több jellemző a listából kevésbé volt hasznos. Habár egy-egy kísérlet lefutása igen gyors volt, mégis óriási számú jellemzővel dolgoztunk, így a kísérletek során nem jutott elegendő számítási kapacitás arra, hogy pontosan meghatározzuk milyen jellemzők azok amik a legjobban segítenek, csak kategória szintjén tudtuk megállapítani, például "a pragmatikai jellemzők

segítettek a legtöbbet" megállapítható, de az nem, hogy ezek közül pontosan melyik jellemzők.
Ezen az adathalmazon csak a szövegbeágyazásokat felhasználva is közel tökéletes eredményeket értünk el, a 10 fold között összesen 1 db hiba volt, ebben az 1 esetben egy emberi szöveget tekintett a modell GPT által írottnak.

4.2. Elírások beillesztése a szövegekbe

A nagyobb adathalmazon, azaz az adathalmazon, amiben 1000db emberi, 999db ChatGPT 4.0-ás, illetve 486db Gemini által generált szövegünk volt, hasonló eredményeket értünk el a gépi kimenetek utófeldolgozása nélkül. Végeztünk kísérleteket azt illetően is, hogy az eredményeket milyen mértékben befolyásolja az, ha véletlenszerűen sokszorosítunk írásjeleket a szövegben: a "." karakterek 10%-a helyett "..." került a szövegbe, a "?" karakterek 10-10%-a helyett "?!" illetve "???", 20%-uk helyett pedig "??". A "!" karakterek 30%-a helyett "!!" karaktereket, illetve 10%-a helyett "!!!" karaktereket helyettesítettünk. Az ilyen módú utófeldolgozás nem változtatott szignifikánsan az eredményeken, azonban egy másikkal már látható változásokat észleltünk az eredményekben. Abban az esetben, ha a validációs halmazban lévő szövegekbe (emberi szövegekbe is) véletlenszerűen szúrtunk elírásokat a korábban részletezett módon, az átlageredmények 10 foldon keresztül a 3. táblázatban leírtak szerint a Macro F1 szerinti sorrendben legjobb 10 eredményt mutatjuk meg.

3. táblázat. Elírásokkal torzított adathalmazon kiértékelt top 10 Macro F1 eredmények

Jellemző-kombinálási módszer	**Használt jellemzők**	**Átlagos Macro F1 pontszám**
attention_on_cat-_and_numerical-_feats	Névelem-felismerés (NER) jellemzők	0.9750
concat	mély morfológiai jellemzők	0.9738
attention_on_cat-_and_numerical-_feats	statisztikai jellemzők	0.9720
weighted_feature-_sum_on-_transformer_cat-_and_numerical-_feats	szintaktikai jellemzők	0.9709
text_only	csak szövegbeágyazások	0.9610
concat	szintaktikai jellemzők	0.9570
gating_on_cat-_and_num_feats-_then_sum	morfológiai jellemzők	0.9561
gating_on_cat-_and_num_feats-_then_sum	helyesírási jellemzők	0.9559
concat	fonetikai jellemzők	0.9557
weighted_feature-_sum_on-_transformer_cat-_and_numerical-_feats	morfológiai jellemzők	0.9557

Hasonló eredményeket kaptunk abban az esetben is, ha csak a gépi szövegekbe szúrtunk elírásokat, ez a 4. táblázatban kerül részletezésre.

4. táblázat. Gépi szövegekben elírásokkal torzított adathalmazon kiértékelt top 10 Macro F1 eredmények

Jellemző-kombinálási módszer	**Használt jellemzők**	**Átlagos Macro F1 pontszám**
attention_on_cat-_and_numerical-_feats	NER jellemzők	0.9736
weighted_feature-_sum_on-_transformer_cat-_and_numerical-_feats	szintaktikai jellemzők	0.9735
concat	mély morfológiai jellemzők	0.9705
attention_on_cat-_and_numerical-_feats	statisztikai jellemzők	0.9653
concat	fonetikai jellemzők	0.9603
weighted_feature-_sum_on-_transformer_cat-_and_numerical-_feats	fonetikai jellemzők	0.9575
weighted_feature-_sum_on-_transformer_cat-_and_numerical-_feats	morfológiai jellemzők	0.9572
gating_on_cat-_and_num_feats-_then_sum	helyesírási jellemzők	0.9572
concat	szintaktikai jellemzők	0.9566
gating_on_cat-_and_num_feats-_then_sum	morfológiai jellemzők	0.9558

A tanítás minden esetben 16-os batch mérettel történt és 3 epoch alatt az elért legjobb eredményt vettük figyelembe. Az epochok során rendszerint keveset változtak az eredmények, sok esetben már az első epochban elérte a legjobb pontosságot. Mindkét táblázat esetén a 10 legjobb konfiguráció olvasható, és a 4. táblázat esetén ha csak szövegbeágyazásokra hagyatkozik a modell, akkor rosszabb teljesítményt érnénk el mind a 10 konfigurációnál. A gépi szövegekben elírásokkal torzított adathalmazon ha csak szövegbeágyazásokra hagyatkozik a modell az átlagos Macro F1 pontszám 0.9549.
A táblázatokban jól megfigyelhetően a legjobb eredményt mindkettő verzió

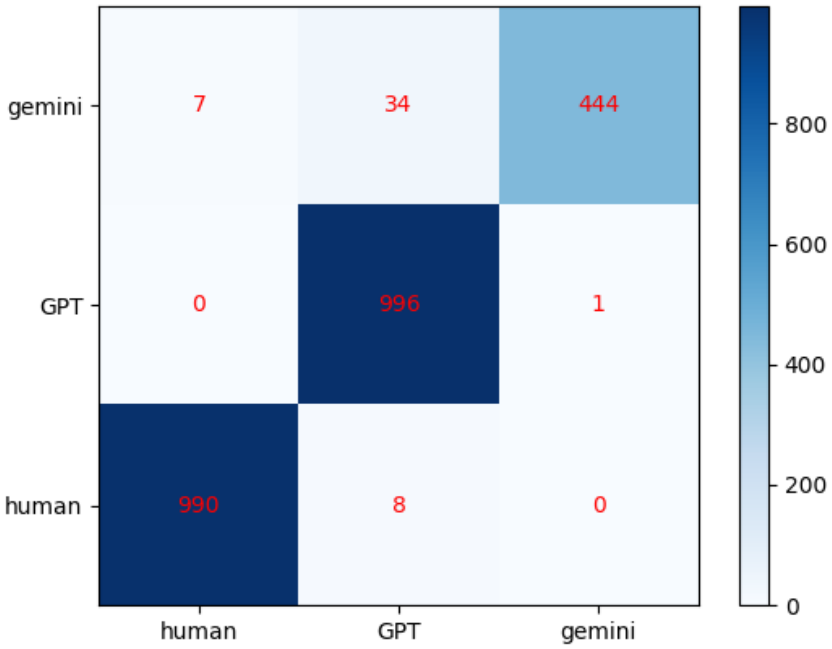
esetén az adta, ha a NER jellemzőket használtuk fel az attention_on_cat_and-_numerical_feats jellemző-kombinálási módszerrel. Ahogy az korábban említve volt, az adat elemzése során megállapítható, hogy a névelemek jelentősen nagyobb arányban fordulnak elő humán szövegekben, mint a gépiekben: a GPT és a Gemini modellek is ritkán írnak konkrét neveket. A mély morfológiai, statisztikai és szintaktikai jellemzők is hasonló eredményeket adnak, bár a többi jellemző sem jelentősen kevésbé pontos. Abban a verzióban, amelyik adathalmazban az emberi szövegekbe is kerültek elírások, csak szövegbeágyazásokat használva is az 5. legjobb eredményt értük el, az említett 4 jellemzőcsoport azonban mind pontosabb eredményhez vezetett.

A két legjobb modell esetén megvizsgáltuk a tévesztési mátrixot is a két elírásos adathalmazban (mindkettő adathalmaznál a legjobb modellt vizsgáltuk). A tévesztési mátrixot kiszámoltuk minden foldban, majd összegeztük. Fontos megjegyezni, hogy minden foldban körülbelül az adathalmaz 10%-a szerepelt a validációs adathalmazban, de nem minden esetben pontosan a 10%-a. Ebből következett, hogy összegezve a 10 foldon keresztül a validációs elemszámokat az eredmény nem pontosan ugyanaz az elemszám, mint ami a teljes adathalmazban is szerepel, hanem marginálisan kevesebb. A tévesztési mátrix ahhoz az adathalmazhoz, melyben a validációs halmaz minden szövegébe kerültek elírások, az 1. ábrán látható.

A tévesztési mátrix alapján levonható következtetés, hogy míg a GPT modell által generált szövegeket sosem predikálta emberinek a modell, a Gemini nagyon ritkán, de sikeresen megtévesztette. Azonban ezen esetek száma is elhanyagolható. A hibák nagyobbik része abból származott, hogy GPT által generált szövegnek tekintette a modell olykor a Gemini által írt szövegeket is, azonban ez szintén nagyon kevés esetben fordult elő.

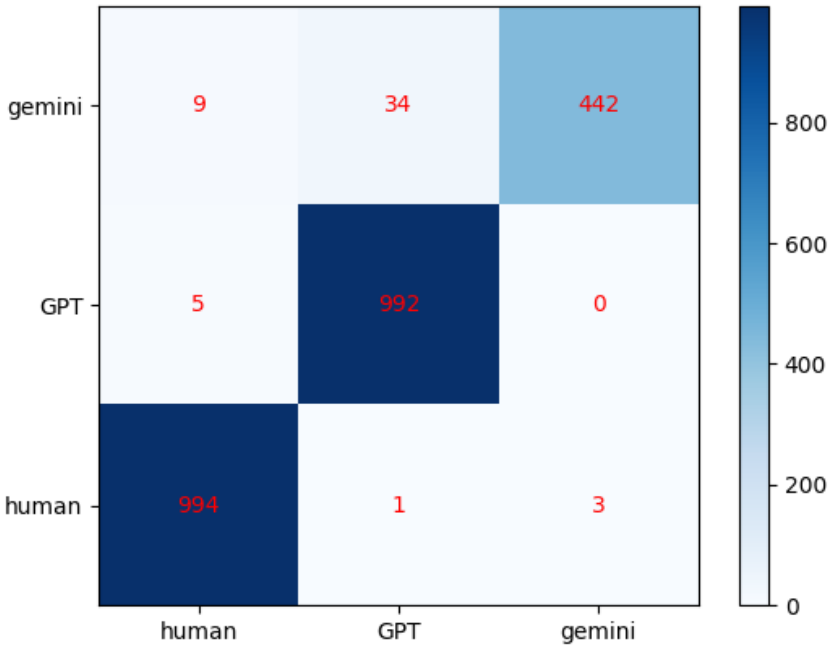
Ugyanilyen módon arra a modellre is megvizsgáltuk a tévesztési mátrixot, melynek tanítása során a validációs halmazban csak a gépi szövegek voltak elírva. Ebben az esetben a 2. ábrán látható tévesztési mátrixot kaptuk.

A mátrixból hasonló eredmények olvashatóak le, mint a másik verzióban is. Láthatóan az ezen a halmazon kiértékelt modell kevésbé pontos, mint a másik, azonban így is ritkán téveszt. Ezen modell esetében 5 esetben a GPT által írt szöveg is emberinek lett prediktálva, a Gemini pedig többször tévesztette



1. ábra. Elírásokkal torzított adathalmaz tévesztési mátrixa. Vízszintes tengely a prediktált címke, függőleges a valódi.

meg, bár csak kevéssel. A leggyakrabban előforduló hiba ebben az esetben is a GPT-nek prediktált Gemini által generált szöveg volt.
A három legjobb modell esetében mindkét elírásos adathalmazra megvizsgáltuk a foldonkénti eredményeket is. A 5. és 6. táblázatban fejtjük ki ezeket az eredményeket bővebben. A 5. táblázatban az 1-es modellhez tartozik a Macro F1 pontszám 1, a 2-eshez a Macro F1 pontszám 2, a 3-ashoz a Macro F1 pontszám 3. Az 5. táblázat is hasonlóan van felépítve. A 5. táblázat esetén az 1-2-3-as modell a 3. táblázat felső 1., 2. illetve 3. modellje, az 6. táblázat esetén pedig a 4. táblázatban lévő modellekre utalunk vissza.



2. ábra. Gépi szövegekben elírásokkal torzított adathalmazon kiértékelt eredmények. Vízszintes tengely a prediktált címke, függőleges a valódi.

5. táblázat. Foldonkénti eredmények a 3 legjobb modell esetén az elírásokkal torzított adathalmazban

Fold	**Macro F1 pontszám 1**	**Macro F1 pontszám 2**	**Macro F1 pontszám 3**
Fold 1	0.9967	0.9967	0.9793
Fold 2	0.9733	0.9700	0.9967
Fold 3	0.9594	0.9803	0.9735
Fold 4	0.9790	0.9843	0.9791
Fold 5	0.9836	0.9539	0.9883
Fold 6	0.9838	0.9671	0.9838
Fold 7	0.9894	0.9785	0.9664
Fold 8	0.9459	0.9948	0.9660
Fold 9	0.9579	0.9579	0.9471
Fold 10	0.9807	0.9540	0.9397

6. táblázat. Foldonkénti eredmények a 3 legjobb modell esetén a gépi szövegekben elírásokkal torzított adathalmazban

Fold	**Macro F1 pontszám 1**	**Macro F1 pontszám 2**	**Macro F1 pontszám 3**
Fold 1	0.9695	0.9946	0.9948
Fold 2	0.9842	0.9732	0.9683
Fold 3	0.9700	0.9853	0.9752
Fold 4	0.9949	0.9900	0.9843
Fold 5	0.9870	0.9775	0.9453
Fold 6	0.9710	0.9953	0.9568
Fold 7	0.9839	0.9631	0.9788
Fold 8	0.9551	0.9400	0.9948
Fold 9	0.9350	0.9323	0.9636
Fold 10	0.9860	0.9840	0.9428

A táblázatokból kiolvasható, hogy mindkét adathalmaz esetén teljesül, hogy kihívás megnevezni egy egyértelmű legjobb konfigurációt: bár az 1-es modell teljesít mindkettő esetben a legjobban átlagosan, mindkét adathalmazban szerepel több fold is, ahol a másik kettő modell közül valamelyik (akár mindkettő) magasabb pontosságot ért el. A két táblázatot összehasonlítva szintén megállapítható, hogy bár a legjobb modell átlageredménye rosszabb azon az adathalmazon, ahol csak a gépi szövegekben voltak elírások, több fold esetén is volt jobb modell az adott foldra mint a másik adathalmazban. Ebből következtethetően bizonyos foldok esetén az emberi szövegek osztályozásában is segített a kisebb mennyiségű elírás, annak ellenére is, hogy a tanítóhalmazban rendszerint jelentősen kevesebb elírás volt emberi szövegekben ezekben a verziókban is, mint a gépi szövegekben (mivel csak a validációs halmaz került utófeldolgozásra), így a validációs halmazban lévő szövegek helyesírás tekintetében közelebb állhattak a modell által látott gépi szövegekhez. Egy lehetséges magyarázata ennek, hogy a szöveg minősége bizonyos foldok esetén fontosabb volt a modellnek a pontos predikcióhoz, mint az, hogy az elírások száma milyen mértékben tükrözi a tanítóhalmazt.

4.3. Független teszthalmazok

Azt a tanított modellünket, mely a legjobban teljesített az utófeldolgozás nélküli validációs halmazokon kiértékeltük a további kisebb, független teszthalmazokon is – ezek szintén megtalálhatóak a mellékletben 6. Kiértékeltük a 28 szövegből álló, GPT által írt szövegen, illetve a 63 darab ParlaMint szövegen is. A ParlaMint szövegek stílusban nagyon különböznek a korábbi emberi szövegeinktől, a parlamenti leiratok miatt jelentősen hivatalosabbak ezek a szövegek. Feltételezhető, hogy a stílus miatt ezek a szövegek sokkal nagyobb hasonlóságot mutatnak a gépi szövegekkel, amik átlagban szintén hivatalosabb formát követnek. A 28 szövegen nagy pontossággal teljesített a modell, 3 kivételével minden szövegről helyesen megállapította, hogy ChatGPT által generált. A 3 hiba esetében Gemini által generáltnak predikálta a szöveget a modell, azaz azokban az esetekben sem tekintette emberinek. A 63 ParlaMint szöveg esetében mind a 63 szövegről helyesen megállapította a modell, hogy ember által írt, azaz 100%-osan pontos volt. Tekintettel arra azonban, hogy a ParlaMint korpusz szövegei jelentősen hosszabbak, végeztünk méréseket olyan módon is, hogy a szövegeket lerövidítettük. Ez annak elkerülése érdekében történt, hogy a modell esetleg minden szövegre, ami nagyon hosszú emberi címkét prediktálna, bár a korábbi kísérleti eredmények nem sugallnak ilyet. Az egyik mérés során az összes szöveg hosszát levágtuk pontosan 4000 karakterre, figyelembe véve, hogy a tanítás/validáció során használt GPT és Gemini szövegek között is voltak ennél hosszabb szövegek is. Egy másik mérés során 3000/3010/3020 ... /3620 karakternél vágtunk, azaz minden szöveg 10 karakterrel hosszabb volt, mint az előző. Mindkét módszerrel megmaradt a 100%-os eredmény, amiből következik hogy nem csupán a szövegek hossza alapján állapította meg a modell, hogy ezek a szövegek emberektől származnak. A rövidített politikai szövegek esetén a modell még abban az esetben is 100%-os pontosságot adott, ha a nyelvi jellemzőket 0-ra állítottuk, és csak a szövegbeágyazásokra hagyatkoztunk (a modell maga nyelvi jellemzőkön is tanult, azaz nem egy text_only modell volt). A gépi írások jellemzően hivatalosabb megfogalmazásúak, mint a hétköznapi emberek írásai, ez alapján viszont megállapítható, hogy a modell nem csupán a szöveg hivatalosabb

megfogalmazása alapján tudja magas pontossággal megkülönböztetni a humán szövegeket a gépi szövegektől; a parlamenti szövegek több névelemmel rendelkeznek, ami szintén jelezheti a szöveg emberi eredetét.
Ezen felül kiértékeltük a modellt a korábban említett, 60 szöveges ChatGPT-s halmazon is, ezen a halmazon jelentősen eltérő eredményeket kapva. A 60 szöveg esetében bár csak 1-ről tudta sikeresen megmondani, hogy ChatGPT írta, további 49-et nem emberi szövegnek prediktált a modell hanem Gemini által írottnak, és csupán a maradék 10 szöveg esetén prediktált emberi szerzőt. A konklúzió, hogy a modellünk stílusban jelentősen eltérő halmazon, rövid szövegek esetén is képes magas pontossággal, 83%-ban megkülönböztetni emberi szerzőt gépi szerzőtől, azonban ilyen halmazon már nem megbízható az az eredménye azt illetően, hogy GPT vagy Gemini által írt-e a szöveg. A jövőben nagyobb tanító adathalmazon tanított modellekkel további kísérletek szükségesek, a chatbotok által változatosabban generált tanítóadattal; az eredmények alapján a változatos tanítóadat hiánya tűnik az elsődleges tévesztési oknak.

5. Összefoglalás, konklúzió

Munkánk során egy olyan módszer kidolgozásán tevékenykedtünk, mely keretén belül megvizsgálható, egy chatbot mennyire hatékonyan tud emberinek tűnő szövegeket generálni, valamint mennyire lehet a szövegek feldolgozásával egy detektáló neurális hálót megtéveszteni. A kutatás során egy egységes kerettel rendelkező promptok segítségével generáltunk adatokat a ChatGPT és Gemini segítségével, melyeket a detektáló hálónk által összehasonlítottunk az emberi szövegekkel. Az egységes promptok mellett független generált teszthalmazokkal is végeztünk méréseket. A szövegek több különböző utófeldolgozáson estek át, így megvizsgálva a torzított adatok detektálásának hatékonyságát. Az adatok előállításában, valamint a felhasználói tapasztalatok összegzésében több munkatársunk is segítséget nyújtott, azonban a kísérletek felépítését, a modellek, valamint nagy részben a szükséges kód implementálását, a kísérletek elvégzését, eredményeik lemérését, illetve a végső következtetések leszűrését a dolgozat két szerzője végezte.

Mind a ChatGPT 4.0, mind a Gemini által generált szöveg hatékonyan detektálható gépi szövegként egyszerűbb, csak szövegre hagyatkozó modellekkel is. A modellek hatékonyságán nem ront az sem, ha az írásjeleket halmozzuk, emberi szerzőket szimulálva. Az azonban ront a modellek hatékonyságán, ha nagy mennyiségű elírást teszünk a szövegekbe, ami mögött az ok feltételezhetően, hogy mind a Gemini modell, mind a GPT modell ritkán vét helyesírási hibákat még akkor is, ha arra külön kérjük. Azonban még az ilyen módon utófeldolgozott adathalmazon is hatékonyan tudnak teljesíteni olyan modellek, melyek a tanítás során csak "tökéletes" gépi szövegeket láttak az emberi szövegek mellett. A NER, mély morfológiai, statisztikai és szintaktikai jellemzők használata mellett a modell pontosabban teljesít mintha csak a szövegbeágyazásokra hagyatkozna. Amennyiben emberi szövegek is gyakran voltak elírva a legjobb eredményt az olvashatósági (alap jellemzőkön) és NER jellemzőkön tanult modell eredményezte 0.9750-ös átlagos Macro F1 pontszámmal, javítva a 0.9610-es átlagos Macro F1 pontszámon, amit csak szövegbeágyazások használata mellett kapnánk. Abban az esetben, ha csak gépi szövegek voltak elírva, szintén a NER és olvashatósági jellemzők adták a legjobb, átlag 0.9736-os Macro F1 pontszámot, ebben az esetben pedig jelentősen rosszabbul teljesített a csak szövegbeágyazásokra alapozó modell, 0.9549-es átlag Macro F1 pontszámot eredményezve.

A konklúzió, hogy utófeldolgozási lépések hiányában a gépi tanuló modelleket mind a ChatGPT 4.0, mind a Gemini modell nehezen tévesztette meg, több különböző promptolási stratégia ellenére is. Egy hatékony utófeldolgozási lépésnek bizonyul elírások beszúrása a szövegbe, azonban a modellek pontosságát ez is csak kis mértékben rontja, különösen, ha azok egyéb nyelvi jellemzőket is figyelembe vesznek a predikcióhoz.

További konklúzió, hogy egyéb promptok által generált szövegek, illetve hivatalos megfogalmazású szövegeken kiértékelve sem állapítható meg romlás a modell teljesítményében, a független teszthalmazokon elért eredményekből következtetve, kivéve a 60 szöveges GPT-s halmazon. Ezen a halmazon is magas pontossággal választ el azonban a modell embert és gépet, mint szerzőt, azonban egy ilyen mértékben eltérő korpuszon, az átlagos szöveghosszhoz képest rövid szövegek esetén már nem feltétlenül képes helyesen megállapítani

a használt modellt. Feltételezhető azonban, hogy változatosabb tanítóadattal, a szövegek több módon történő elő- és utófeldolgozásával javítható a teljesítmény.

6. További irányok, kitekintés

Kísérleteinkkel azt vizsgáltuk, hogy a modellünk képes-e teljes szövegekről eldönteni, azokat ember, vagy mesterséges intelligencia írta-e. Kutatásunk folytatásaként vizsgálatokat végezhetnénk hibrid szövegekkel is; olyan szövegekkel, melyek vegyesen tartalmaznak emberi, illetve mesterséges intelligencia eredetű szövegrészleteket is. Érdemes lehet megvizsgálni, a szövegrészek keverése milyen módon befolyásolná a detektáló hálónkat: a hibrid szöveg 1-1 arányú emberi-gépi szövegtartalmazása esetén hasonló eredmények születnének-e, mint például az 1-4, vagy 4-1 arány esetében; ha a hibrid szövegben egy blokkban lenne az emberi szövegrész, majd előtte/utána a neurális háló szövegrésze, illetve, teljesen heterogénen keverve a szöveget – akár bekezdésenként, de akár sokkal megengedőbb módon is keverhetjük a szövegrészeket, például gondolatonként, mondatonként –, akkor milyen eredményeket tudnánk elérni. A hibrid adatok bevezetésével a kiértékelésünkön is módosíthatunk, ami maga után vonná a kiértékelő modellünk működésének és számosságának potenciális változását. Bevezetésre kerülhetne például egy ötös (nagyon nem valószínű, nem valószínű, nem lehet megállapítani / kétes, valószínű, nagyon valószínű) osztályozás – ezt a felosztást gyakran használják a hármas negatív-semleges-pozitív mellett –, mely a háló konfidenciáján alapulna [5] – továbbá, ezt a kiértékelési metrikát összevonhatjuk a jelenlegi hármas osztályozással, így megkapva mind a három osztály konfidenciáját; ezen felül magukat a konfidenciákat is megvizsgálhatjuk az egyes predikcióknak, így potenciálisan korrelációt vonva a gépi szöveg (százalékos) tartalmával.
A ChatGPT 4.0-val és a Google Gemini modelljével elért eredményeket érdemes lehet összehasonlítani más generatív neurális hálókkal is: például a két háló elődjével, vagyis a ChatGPT 3.5-tel és a Google Barddal (utóbbi esetében lehetséges, hogy nem létezik már különálló egységként egy régebbi modell verzió, így ezt is meg kell előzetesen vizsgálni) lenne érdemes kísérleteket tenni

– így betekintést kaphatnánk arról, hogy ezen a területen milyen fejlődésen estek át a technológiák. A régebbi modellek mellett érdemes lehet a többi, potenciálisan erős neurális hálókkal is kísérleteket végezni – a jelenleg nagy népszerűségnek örvendő chatbotok [34] közül a következő modelleket lehet érdemes megvizsgálni, például:

- Claude 2: erőssége, hogy nagyon sokáig vissza tud emlékezni a beszélgetés korábbi részleteire
- Microsoft Bing AI: keresőfunkciójának köszönhetően egy széles körben ismert modell
- 60 db vegyes, ChatGPT 3.0 és ChatGPT 4.0 által generált szöveg
- Llama 2 és 3: a Meta saját, nyílt licencű keretrendszerrel rendelkező AI szolgáltatása. A Llama 2 már 2023 februárja óta működik, míg a Llama 3 frissen jelent meg [14]. Itt is érdekes lehet összevetni a két modell közötti teljesítményt
- Pi: személyes intelligencia - personal intelligence rövidítése, az Inflection AI [13] saját fejlesztésű, az egyéni felhasználásra optimalizált mesterséges intelligenciája

A felsorolt neurális hálók között csak olyan szerepel, melyek természetesnyelvű kommunikációra lettek tanítva; például a GitHub Copilot azért nem szerepel az elsődlegesen megvizsgálandó mesterséges intelligenciák között, mivel elsősorban programkód-természetesnyelv interakciókra tanították.
Az utófeldolgozási metodikáinkat, eljárásainkat is bővíthetnénk: jelenleg főleg az elírásokat tettünk a szövegbe. Ennek fordítottját, az emberi szövegek helyesírás-javított változataival való kísérleteket is érdemes lenne megnézni, hogy magasabb eséllyel prediktálna-e a modell GPT/Gemini címkéket emberi szövegekre. Ezen felül más jellegű perturbációkat [39] is alkalmazhatnánk az adathalmazokon: a bekezdések számát, hosszát is lehetne módosítani, illetve a modellt a szemantikai szempontból felcserélhető szövegrészek különböző sorrendkombinációiból alkotott szövegeken is érdemes lenne tesztelni.
A tartalom, struktúra, valamint beszédstílus alapján specifikus szövegekkel is

releváns lehet kísérleteket végezni: esszék, orvosi diagnózisok, gyermekek által fogalmazott szövegek, ADHD-val diagnosztizált emberek írásai.

Köszönetnyilvánítás

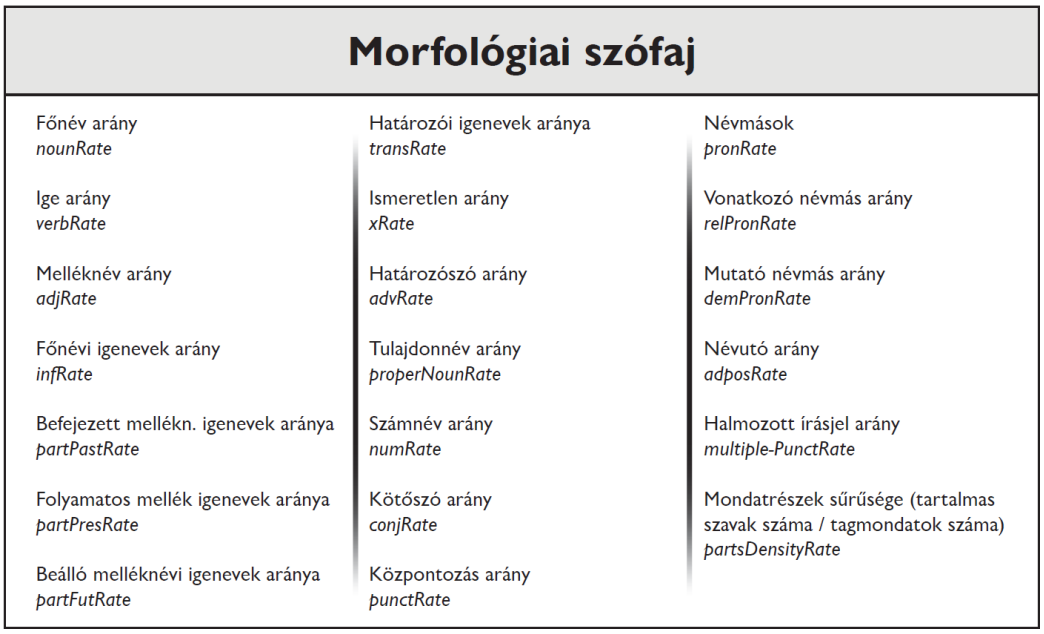
A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A kutatás a TKP2021-NVA-09 projekt támogatásával készült. A TKP2021-NVA-09 számú projekt Magyarország Innovációs és Technológiai Minisztériumának támogatásával valósult meg a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból, a TKP2021-NVA támogatási finanszírozási keret alapján.
A kutatás során többen segítettek, nekik szeretnénk megköszönni külön a munkát. Köszönjük Balogh Etelének a személyiség leírásokat a promptokhoz, illetve a nagyobb (több mint 2000 szöveges) adathalmazt illetően az alap promptot is, illetve a különböző tanácsokat, teszteléseket is. Ezen kívül köszönjük Kálmán Gabriella nyelvésznek a Gemini adathalmaz legenerálását, illetve a meglátásait. Szintén köszönetet nyilvánítunk Főző Eszter nyelvészszakértőnek további ChatGPT-s tesztadataiért, meglátásaiért.

Melléklet

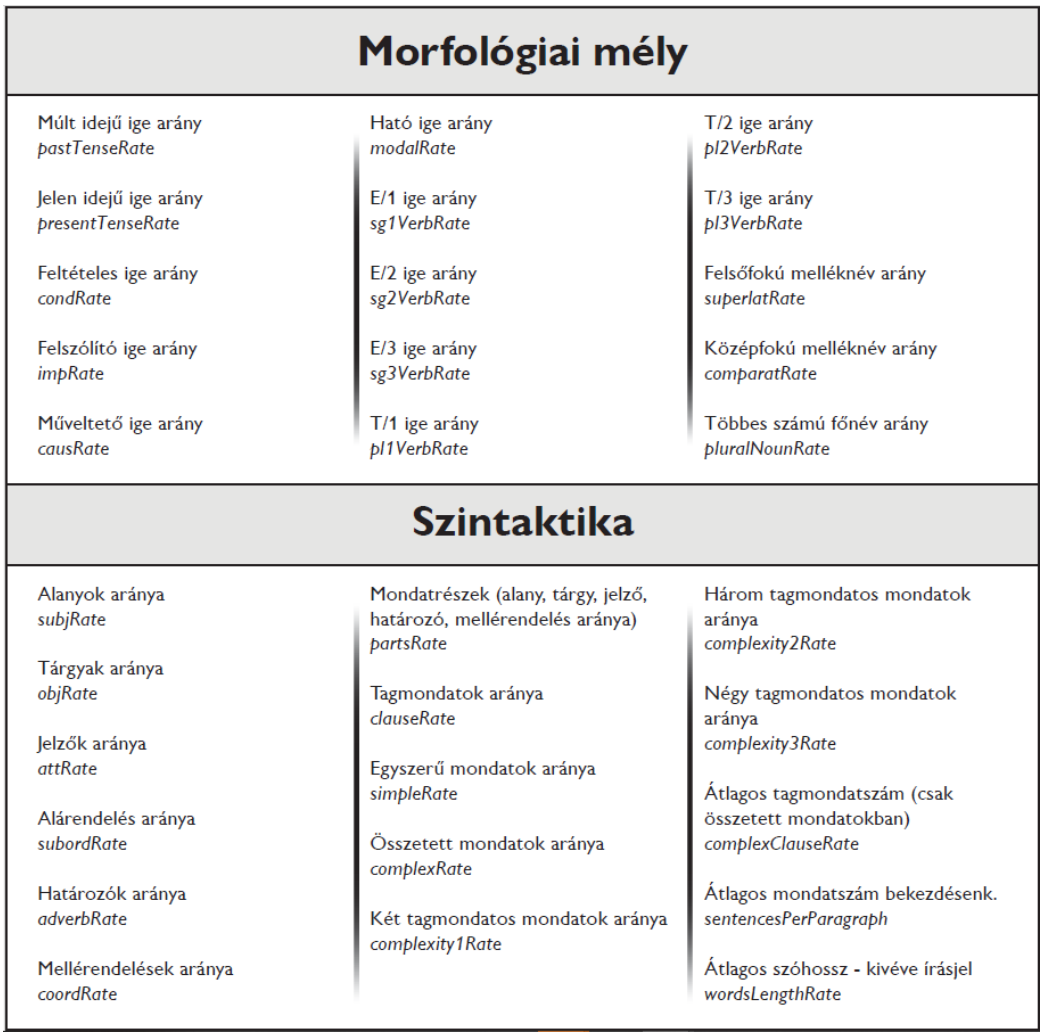
A kísérletekhez használt adatok, promptok, személyiségek a tdk_gepelted_vagy_gepelted GitHub repozitóriumában[1] érhetőek el.

[1] https://github.com/TDKGitHubUser/tdk_gepelted_vagy_gepelted

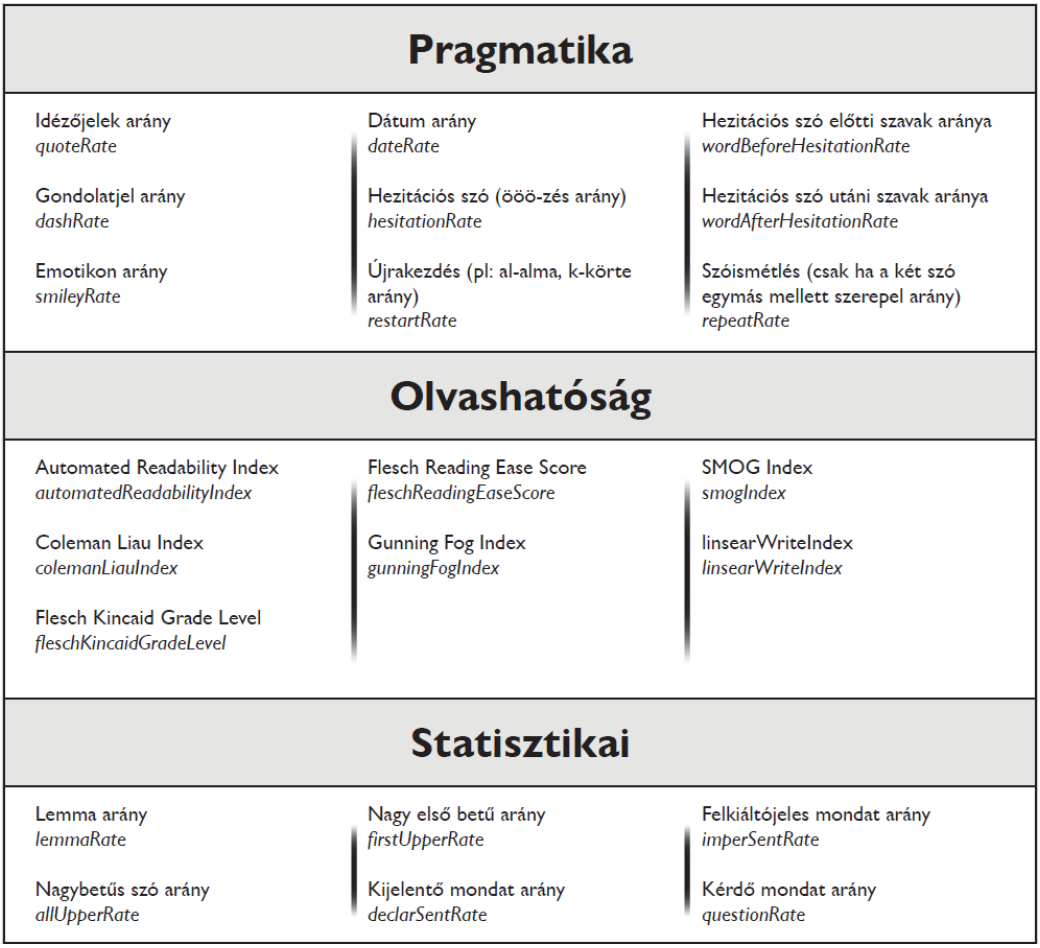
Nyelvi jellemzők



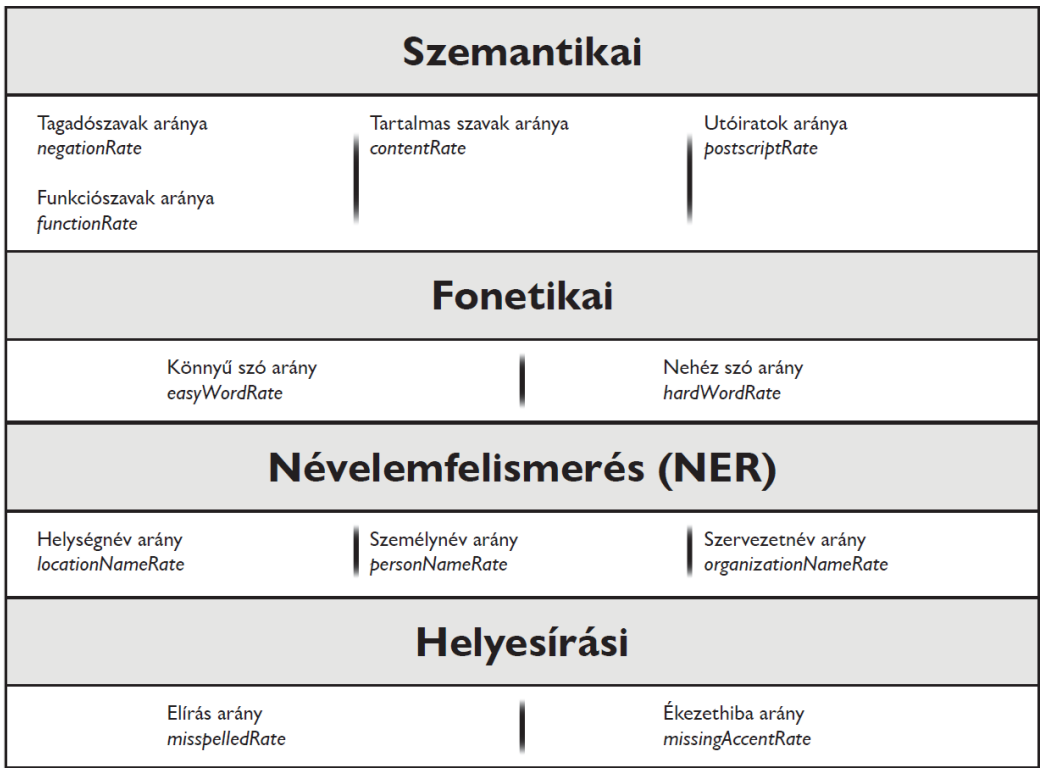
3. ábra. Morfológiai jellemzőlista



4. ábra. Morfológiai mély, szintaktika jellemzőlista



5. ábra. Pragmatika, olvashatóság, statisztikai jellemzőlista



6. ábra. Szemantikai, fonetikai, NER, helyesírási jellemzőlista

Felhasznált irodalom

[1] Linnea Ahlgren. *Google's Gemini AI won't be available in Europe — for now — thenextweb.com.* https://thenextweb.com/news/google-gemini-ai-unavailable-europe-uk. [Accessed 22-04-2024].

[2] Anirban Chakraborty és tsai. "A survey on adversarial attacks and defences". *CAAI Transactions on Intelligence Technology* 6.1 (2021. márc.), 25–45. old. ISSN: 2468-2322. DOI: 10.1049/cit2.12028. URL: http://dx.doi.org/10.1049/cit2.12028.

[3] *ChatGPT and Bard can generate Windows keys, but there's a catch — windowscentral.com.* https://www.windowscentral.com/software-apps/windows-11/chatgpt-and-bard-can-generate-windows-keys-but-theres-a-catch. [Accessed 22-04-2024].

[4] K. R. Chowdhary. "Natural Language Processing". *Fundamentals of Artificial Intelligence*. New Delhi: Springer India, 2020, 603–649. old. ISBN: 978-81-322-3972-7. DOI: 10.1007/978-81-322-3972-7_19. URL: https://doi.org/10.1007/978-81-322-3972-7_19.

[5] Ahmed M. Elkhatat, Khaled Elsaid és Saeed Almeer. "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text". *International Journal for Educational Integrity* 19.1 (2023), 17. old. ISSN: 1833-2595. DOI: 10.1007/s40979-023-00140-5. URL: https://doi.org/10.1007/s40979-023-00140-5.

[6] Tomaž Erjavec és tsai. *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0*. Slovenian language resource repository CLARIN.SI. 2023. URL: http://hdl.handle.net/11356/1859.

[7] Louie Giray. "Prompt Engineering with ChatGPT: A Guide for Academic Writers". *Annals of Biomedical Engineering* 51.12 (2023), 2629–2633. old. ISSN: 1573-9686. DOI: 10.1007/s10439-023-03272-4. URL: https://doi.org/10.1007/s10439-023-03272-4.

[8] G. M. Gritsay és tsai. “Artificially Generated Text Fragments Search in Academic Documents”. *Doklady Mathematics* 108.2 (2023), S434–S442. ISSN: 1531-8362. DOI: 10.1134/S1064562423701211. URL: https://doi.org/10.1134/S1064562423701211.

[9] Ken Gu és Akshay Budhkar. “A Package for Learning on Tabular and Text Data with Transformers”. *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. Szerk. Amir Zadeh és tsai. Mexico City, Mexico: Association for Computational Linguistics, 2021. jún., 69–73. old. DOI: 10.18653/v1/2021.maiworkshop-1.10. URL: https://aclanthology.org/2021.maiworkshop-1.10.

[10] Riccardo Guidotti és tsai. “A Survey of Methods for Explaining Black Box Models”. *ACM Comput. Surv.* 51.5 (2018. aug.). ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009.

[11] Hani Hagras. “Toward Human-Understandable, Explainable AI”. *Computer* 51.9 (2018), 28–36. old. DOI: 10.1109/MC.2018.3620965.

[12] Levente Halmosi és Mark Jelasity. *On Evaluating the Adversarial Robustness of Semantic Segmentation Models*. 2023. arXiv: 2306.14217 [cs.CV].

[13] *Inflection — inflection.ai*. https://inflection.ai/. [Accessed 22-04-2024].

[14] *Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com*. https://ai.meta.com/blog/meta-llama-3/. [Accessed 22-04-2024].

[15] András Kicsi és tsai. “Computer-Aided Forensic Authorship Identification in Criminology”. *Computational Science and Its Applications – ICCSA 2022 Workshops*. Szerk. Osvaldo Gervasi és tsai. Cham: Springer International Publishing, 2022, 576–592. old. ISBN: 978-3-031-10548-7.

[16] Yoonsu Kim és tsai. “Understanding Users’ Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level”. *Proceedings of the 29th International Conference on Intelligent User*

Interfaces. IUI '24. ACM, 2024. márc. DOI: 10.1145/3640543.3645148. URL: http://dx.doi.org/10.1145/3640543.3645148.

[17] Nikita Kitaev, Steven Cao és Dan Klein. "Multilingual Constituency Parsing with Self-Attention and Pre-Training". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. júl., 3499–3505. old. DOI: 10.18653/v1/P19-1340. URL: https://www.aclweb.org/anthology/P19-1340.

[18] Nikita Kitaev és Dan Klein. "Constituency Parsing with a Self-Attentive Encoder". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. júl., 2676–2686. old. DOI: 10.18653/v1/P18-1249. URL: https://www.aclweb.org/anthology/P18-1249.

[19] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". 14 (2001. márc.).

[20] Németh László. *Hunspell*. [Accessed 29-04-2024]. URL: https://hunspell.github.io/.

[21] Klas Leino, Zifan Wang és Matt Fredrikson. "Globally-Robust Neural Networks". *Proceedings of the 38th International Conference on Machine Learning*. Szerk. Marina Meila és Tong Zhang. 139. köt. Proceedings of Machine Learning Research. PMLR, 2021. júl., 6212–6222. old. URL: https://proceedings.mlr.press/v139/leino21a.html.

[22] Miriam Meyerhoff. *Introducing Sociolinguistics*. Routledge, 2018. aug. ISBN: 9780429507922. DOI: 10.4324/9780429507922. URL: http://dx.doi.org/10.4324/9780429507922.

[23] Prakash M Nadkarni, Lucila Ohno-Machado és Wendy W Chapman. "Natural language processing: an introduction". *Journal of the American Medical Informatics Association* 18.5 (2011. szept.), 544–551. old. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000464. eprint: https://

academic.oup.com/jamia/article-pdf/18/5/544/5962687/18-5-544.pdf. URL: https://doi.org/10.1136/amiajnl-2011-000464.

[24] Dávid Márk Nemeskey. "Introducing huBERT". *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*. Szeged, 2021, TBA.

[25] OpenAI és tsai. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL].

[26] György Orosz és tsai. *HuSpaCy: an industrial-strength Hungarian natural language processing toolkit*. 2022. arXiv: 2201.01956 [cs.CL].

[27] *Papers with Code - Leaky ReLU Explained — paperswithcode.com*. https://paperswithcode.com/method/leaky-relu. [Accessed 22-04-2024].

[28] Wasifur Rahman és tsai. "Integrating Multimodal Information in Large Pretrained Transformers". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Szerk. Dan Jurafsky és tsai. Online: Association for Computational Linguistics, 2020. júl., 2359–2369. old. DOI: 10.18653/v1/2020.acl-main.214. URL: https://aclanthology.org/2020.acl-main.214.

[29] Partha Pratim Ray. "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope". *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154. old. ISSN: 2667-3452. DOI: https://doi.org/10.1016/j.iotcps.2023.04.003. URL: https://www.sciencedirect.com/science/article/pii/S266734522300024X.

[30] Pablo Rivas és tsai. "Application-Agnostic Chatbot Deployment Considerations: A Case Study". *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. 2019, 361–365. old. DOI: 10.1109/CSCI49370.2019.00070.

[31] Adi Robertson. *Google apologizes for "missing the mark" after Gemini generated racially diverse Nazis — theverge.com*. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical. [Accessed 22-04-2024].

[32] Vinu Sankar Sadasivan és tsai. *Can AI-Generated Text be Reliably Detected?* 2024. arXiv: 2303.11156 [cs.CL].

[33] Gemini Team és tsai. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. arXiv: 2312.11805 [cs.CL].

[34] *The best AI chatbots in 2024 | Zapier — zapier.com*. https://zapier.com/blog/best-ai-chatbot/.

[35] Dániel Varga és tsai. "Parallel corpora for medium density languages". *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* 292 (2007), 247. old.

[36] *Watch how ChatGPT is tricked into generating Windows 95 keys — pcworld.com*. https://www.pcworld.com/article/1681090/watch-how-chatgpt-is-tricked-into-generating-windows-95-keys.html. [Accessed 22-04-2024].

[37] Debora Weber-Wulff és tsai. "Testing of detection tools for AI-generated text". *International Journal for Educational Integrity* 19.1 (2023), 26. old. ISSN: 1833-2595.

[38] Jiliang Zhang és Chen Li. "Adversarial Examples: Opportunities and Challenges". *IEEE Transactions on Neural Networks and Learning Systems* 31.7 (2020), 2578–2593. old. DOI: 10.1109/TNNLS.2019.2933524.

[39] Ying Zhou, Ben He és Le Sun. *Humanizing Machine-Generated Content: Evading AI-Text Detection through Adversarial Attack*. 2024. arXiv: 2404.01907 [cs.CL].