Eötvös Loránd Tudományegyetem

Társadalomtudományi kar

MESTERKÉPZÉS SZAKDOLGOZAT

Mesterséges intelligencia által generált és ember által írt kutatási absztraktok klasszifikációja

Konzulens:

Buda Jakab Máté

Készítette:

Makláry János Zsolt

E4BNW4

Survey statisztika és adatanalitika szak

2025, április

EREDETISÉGNYILATKOZAT

(Kitöltés után a szakdolgozat/diplomamunka részét képezi.)

Alulírott ..... Makláry János Zsolt..... *[a hallgató neve]* az ELTE TáTK

.....survey statisztika és adatanalitika..... *[a hallgató szakja]* hallgatója

fegyelmi felelősségem tudatában nyilatkozom, és aláírásommal igazolom, hogy a(z)

..... Mesterséges intelligencia által generált és ember által írt kutatási absztraktok klasszifikációja..... *[a szakdolgozat/diplomamunka címe]*

című szakdolgozat/diplomamunka **saját, önálló szellemi munkám,** az abban hivatkozott, nyomtatott és elektronikus szakirodalom felhasználása a szerzői jogok szabályainak megfelelően történt.

Tudomásul veszem, hogy szakdolgozat/diplomamunka esetén plágiumnak számít:

- a szó szerinti idézet vagy fordítás közlése idézőjel és hivatkozás megjelölése nélkül;
- a tartalmi idézet hivatkozás megjelölése nélkül;
- más publikált gondolatainak saját gondolatként való feltüntetése.

Alulírott kijelentem, hogy a plágium fogalmát, valamint a HKR 74/A–C. §-aiban foglalt

rendelkezéseket megismertem, és tudomásul veszem, hogy

- a szakdolgozatom/diplomamunkám szövege a benyújtása után plágiumellenőrző szoftverrelvizsgálható,
- plágium esetén szakdolgozatom/diplomamunkám értékelés nélkül visszautasításra kerül,
- plágiumvétség esetén fegyelmi eljárás indítható.

Budapest/Szombathely, 2025. április 12.

.....Makláry János Zsolt.....

a hallgató aláírása

Kivonat

Dolgozatomban azt vizsgálom, hogy klasszikus gépi tanulási algoritmusok milyen pontossággal képesek mesterséges intelligencia által generált tartalmat felismerni egy modern, transzformer alapú detektorhoz viszonyítva, továbbá az AI akadémiai közegre gyakorolt hatásaival is foglalkozom.

Kulcsszavak

akadémia, plágium, mesterséges intelligencia, szöveggenerálás, klasszifikáció, detekció

Tartalomjegyzék

I. Bevezetés

I.1 Kutatási kérdés

Az utóbbi években a mesterséges intelligencia fejlődésének köszönhetően olyan kifinomult szoftverek jöttek létre, amelyek képesek olyan magas színvonalon szöveget generálni, hogy az nehezen különböztethető meg az ember által írt tartalmaktól. Szakdolgozatom tágabb kontextusában azt vizsgálom, hogy milyen hatása van ennek a jelenségnek az akadémiai közegre, konkrét kutatási kérdésem pedig a következő:

> *Milyen pontossággal képesek a hagyományos klasszifikációs algoritmusok detektálni mesterséges intelligencia által generált szövegeket modern, transzformer alapú detektorokhoz képest?*

I.2 A mesterséges intelligencia köztudatba férkőzése

Az elmúlt néhány év technológiai fejlődésének középpontjában vitathatatlanul a mesterséges intelligencia állt. A különböző AI[1]-on alapuló applikációk széles körben elterjedté váltak, személyes és munkahelyi tevékenységhez kapcsolódó használatuk a mindennapi élet részévé vált. Szignifikáns szerepet játszott ebben az áttörésben az amerikai OpenAI vállalat, 2022 novemberében megjelentetett ChatGPT névre keresztelt chatbot-juk tekinthető ugyanis az első igazán nagy tömeget megszólító mesterséges intelligencia alapú alkalmazásnak. A Reuters[2] (2023) értesülései szerint a ChatGPT 3.0 2023 januárjára már 100 millió legalább havi szintű aktív felhasználóval rendelkezett, és ezzel a történelem leggyorsabban növekedő applikációjává vált. Viszonyításképpen ugyanekkora méretű felhasználóbázis elérésére a TikTok-nak kilenc hónapra, az Instagram-nak két és fél évre volt szüksége (Reuters, 2023). A Salesforce [3] (2025) frissen végzett kutatása szerint, amelyben több mint 4000 teljes munkaidőben dolgozó alanyt kérdeztek mesterséges intelligenciával kapcsolatos attitűdjeikről és használati szokásaikról, a szellemi munkát végzők 61%-a használ jelenleg is, vagy tervez használni a közeljövőben mesterséges intelligencia alapú megoldásokat munkája során. Ezzel

[1] artificial intelligence – a későbbiekben használom ezt a rövidítést
[2] Reuters. (2023). ChatGPT sets record for fastest-growing user base: Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[3] Salesforce. (2025). Top generative AI statistics for 2025. Salesforce News. https://www.salesforce.com/news/stories/generative-ai-statistics/

egybehangzó eredményeket hozott McKinsey[4] 2024-es globális online felmérése is, miszerint a megkérdezett vállalatvezetők 65%-a integrált valamilyen AI alapú technológiát céges működésükbe (McKinsey & Company, 2024).

A terület rohamos fejlődését serkenti az is, hogy világszerte számos nagyvállalat kezdett mesterséges intelligencia alapú megoldásokat fejleszteni, ami egyre élesedő versenyhelyzetet teremtett az iparágban. Az innovatív fejlesztések tekintetében az OpenAI mellett kulcsszereplőnek tekinthető a teljesség igénye nélkül a Meta (Llama), a Google DeepMind (Gemini), a Microsoft (Microsoft Copilot) és az AmazonWebServices (Titan) is (Credence Research, 2025[5]), de új szereplők is feltűnnek a piacon, mint például a 2023-ban Kínában alapított DeepSeek, melynek legfrissebb, 2025 januárjában megjelenő nagy nyelvi modellje (DeepSeek R1) alaposan felborzolta kedélyeket az AI szektorban (BBC News, 2023[6]). A globális nagyhatalmak között létrejövő „AI verseny" tovább serkentheti az amúgy sem lassú mértékű technológiai fejlődést.

Látható, hogy a mesterséges intelligencia térnyerése a 2020-as évek közepére olyan mértékűvé vált, hogy indokolt lehet annak társadalmi hatásaival is foglalkozni. Mahajan (2023) szerint tagadhatatlan, hogy a mesterséges intelligenciához hasonló mértékű technológiai újítások a társadalom számára lehetőségeket és veszélyeket egyaránt hordoznak. Egyfelől számos területen alkalmazhatók olyan AI alapú megoldások, melyek társadalmilag pozitív hatással bírhatnak. Ilyen terület az orvostudomány, ahol különböző mesterséges intelligencia alapú módszerek segíthetnek például pontosabb diagnózis felállításában, vagy személyre szabott terápia megválasztásában, de az oktatás fejlesztésében is nagy szerepet játszhatnak AI alapú technológiák, melyek területtől függetlenül serkentő hatással lehetnek a hatékonyságra és a produktivitásra (Mahajan, 2023). Másfelől viszont problémás kérdések is felmerülnek a témakör kapcsán: milyen hatással lesz a mesterséges intelligencia fejlődése a munkaerőpiacra, veszélyeztetheti-e az AI emberek megélhetését? Mennyire megbízható módon működnek a különböző AI alapú technológiák, hogyan kezelik az adott esetben privát információt? Milyen hatással lehetnek különböző nagy nyelvi modellek a hamis információ terjedésére? Milyen

[4] McKinsey & Company. (2024). The state of AI 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[5] Credence Research. (2025). Large language model market. https://www.credenceresearch.com/report/large-language-model-market

[6] BBC News. (2023). ChatGPT sets record for fastest-growing user base. https://www.bbc.com/news/articles/c5yv5976z9po

etikai aggályok merülnek fel AI által generált tartalom használatát illetően? Összességében tehát nagy potenciál rejlik a mesterséges intelligencia fejlődésében társadalmunk szempontjából, de elengedhetetlen valamilyen körültekintő jogi és etikai szabályrendszer kialakítása.

I.3 AI az akadémiában: lehetőségek, kihívások

A mesterséges intelligencia fejlődése ahogy oly sok területre, a tudományos, akadémiai közegre is komoly hatással lehet (Miao és munkatársai, 2024). Megítélése azonban kettős: egyfelől izgalmas lehetőségeket tartogat, melyekkel mind a kutató-, mind az oktatómunka megkönnyíthető és hatékonyabbá tehető, másfelől viszont komoly etikai kérdések is felmerülnek használatát illetően (Meyer és munkatársai, 2023).

Lehetséges használati módok

Kitűnő szövegösszefoglalási készségüknek köszönhetően nagyszerűen alkalmazhatók nagy nyelvi modellek a kutatómunka kezdeti fázisában, a szakirodalom áttekintése során. Az eseteként akár több tíz oldalas tudományos cikkek vagy hosszas könyvfejezetek, adott esetben egész könyvek áttekintése és értelmezése kifejezetten idő és energiaigényes feladat lehet, a kutatómunka jelentős részét teheti ki. Mesterséges intelligencia segítségével egyszerre akár több kutatás is áttekinthető, eredményeik együttesen értelmezhetők lehetnek (Guo és munkatársai, 2023). Delgado-Chaves és munkatársai (2025) azt vizsgálták, hogy szisztematikus irodalomáttekintés során mennyire jól alkalmazhatók különböző nagy nyelvi modellek, amelyek a kutatásban cím és absztrakt alapján vizsgálták felül az egyes tanulmányokat. Azt találták, hogy a vizsgált LLM[7] modellek nagy mértékben le tudják csökkenteni az irodalom áttekintésével járó munka mennyiségét, ugyanis rendkívül gyorsan képesek eldönteni, hogy egy adott tanulmány releváns lehet-e az eredeti kutatás szempontjából. A szerzők hozzáteszik azonban, hogy az ember által ráfordított idő nem vonható ki teljes mértékben a folyamatból, ugyanis az nagy nyelvi modellek által elvégzett munka mindenképpen szakértői ellenőrzésre szorul (Delgado-Chaves és munkatársai, 2025). Ennek oka az lehet, hogy bizonyos nagy nyelvi modellek, mint például a ChatGPT, hajlamosak lehetnek téves információt is közölni, máshogy szólva „hallucinálni”. Meyer (2023) ennek okán azt javasolja, hogy a szöveget generáltató

[7] large language model – a későbbiekben használom ezt a rövidítést

személy saját munkájaként vállaljon felelősséget a tartalom hitelességéért. Egyéb iránt többek között a hallucináció problémájára próbálnak megoldást nyújtani azok a speciális LLM-ek, amelyek kifejezetten tudományos forrásokat alapul véve válaszolják meg a felhasználó kérdéseit (Meyer és munkatársai, 2023), ilyen modell például a perplexity.ai.

Segítségül hívható mesterséges intelligencia akkor is, amikor egy kutatás módszertanát szeretnénk megválasztani. Salvagno és szerzőtársai (2023) tanulmányukban kifejtik, hogy csupán nyers adatok alapján a ChatGPT képes lehet asszisztálni egy cikk módszertani szekciójának megírásában: indokolni tudja a mintanagyság megválasztását és képes bemutatni az alkalmazandó adatelemzési technikákat. Mondal és munkatársai (2024) kutatásukban azt vizsgálták, hogy ingyenesen elérhető nagy nyelvi modellek (ChatGPT 3.5, Microsoft Bing, Google Bard) mennyire képesek helyes statisztikai metódust javasolni bizonyos fiktív kutatási szcenáriók esetén. Minden ilyen kutatási tervhez az adott téma szakértőiétől kértek javaslatot statisztikai elemzési módszerre, majd ezeket összevetették a nagy nyelvi modellek által javasolt módszerekkel. Minden LLM esetében elmondható, hogy az egyezés legalább 75%-os volt, a legjobb eredményt a Bing érte el (96,3%). Az eredményekből az a következtetés vonható le, hogy a kutatás tervezési fázisában döntéstámogató eszközként hatékony segítséget nyújthatnak az említett nagy nyelvi modellek, fontos azonban hozzátenni, hogy nem helyettesítik az emberi szakértelmet (Mondal és munkatársail, 2024)

A 2022 novemberében megjelenő ChatGPT 3 nevű nagy nyelvi modell azért számított igazán áttörőnek, mert elődeivel ellentétben úgy volt megtervezve, hogy különböző természetesnyelv feldolgozáshoz kapcsolódó feladatok széles körére alkalmas legyen. Ilyen feladatok például a szövegösszegzés, fordítás, kreatív írás, hosszabb vagy rövidebb szöveggenerálás (Eke, 2023). Az akadémiai kutatómunka talán legfontosabb részfeladata a kutatás bemutatása, a tudományos publikáció szövegének megírása. Generatív jellegükből adódóan ebben a fázisban is kitűnően alkalmazhatók különböző nagy nyelvi modellek, sőt, talán itt mutatkozhat meg leginkább a bennük rejlő potenciál. Salvagno és szerzőtársai (2023) szerint – akik a ChatGPT akadémiai alkalmazhatóságát tanulmányozták – cikkek vázlatainak megírására, címadásra, szövegformázásra, nyelvtani hibák kiszűrésére, bonyolultan megfogalmazott szövegrészek leegyszerűsítésére vagy akár az absztrakt megalkotására is alkalmasak lehetnek nagy nyelvi modellek. Sikander és munkatársai (2023) kutatásukban mesterséges intelligencia által generált és ember által írt kutatási absztraktokat hasonlítottak össze az alábbi módon: a téma

szakértőit kérték meg, hogy értékeljenek összesen 18 szövegpárt (ember írta – AI generálta) publikálhatóság, olvashatóság és tartalmi minőség alapján. Az eredmények azt mutatták, hogy nem minősültek alsóbbrendűnek a mesterséges intelligencia által készített szövegek: nem volt szignifikáns különbség a publikálhatóságot és a tartalmi minőséget illetően, az olvashatóság tekintetében pedig szignifikánsan magasabb értékelést kaptak a nem ember által írt absztraktok. Továbbá, az értékelést végző személyek 59%-a mesterséges intelligencia által generált szövegeket részesítette előnyben, a legfőbb különbségeket abban látták, hogy az AI által készített szövegek „szebben" voltak fogalmazva és magukkal ragadóbbak voltak (a szerző itt a „better flow" kifejezést használja), míg az ember írta bevezetők relevánsabb háttérinformációt tartalmaztak és jobb struktúrával rendelkeztek (Sikander és munkatársai, 2023). Összeségében tehát az látható, hogy hatékonyan alkalmazhatók nagy nyelvi modellek az akadémiai szövegalkotás különböző területein, de a megfelelő minőség és akkurátusság elérése érdekében az emberi szakértelem a munka ezen fázisaiból sem hagyható ki.

Dilemmák – veszélyben az akadémiai integritás?

Egy nemzetközi akadémiai integritással foglalkozó szervezet (Internation Center for Academic Integrity) hat alapvető érték[8] iránti elköteleződésként definiálja az akadémiai integritást. Az értékek a következők: őszinteség, bizalom, igazságosság, tisztelet, felelősség és bátorság. Ezen értékek nélkül a kutatók, oktatók és diákok munkája hitelességét és értékét veszítheti. Azzal, hogy egy kutató vagy diák tanulmányában, vagy dolgozatában mesterséges intelligencia által generált szöveget használ, megsérti az akadémiai integritás alapvető értékeit (Eke, 2023). Ehhez hasonló aggályokat sok más szerző is megfogalmaz, Hua és munkatársai (2024) szerint a ChatGPT forradalmi újításai serkentően hathatnak a plagizálásra, veszélybe sodorba ezzel az akadémiai integritást. Miao és szerzőtársai (2024) szerint ideális esetben a mesterséges intelligencia abban segíthetné a kutatókat, hogy a monoton feladatok helyett formabontó ötleteikre és izgalmas problémák megoldására fókuszálhassanak. A valóság azonban merőben eltér ettől. Egyre több olyan kutatás számol be olyan esetekről, amelyekben szakmai elbíráláson átesett cikkekről derül ki, hogy tartalmuk nem teljes egészében emberi munka szüleménye (Miao és munkatársai, 2024). Már a 2020-21-es tanévre vonatkozóan is fellelhetők olyan adatok, amelyek azt bizonyítják, hogy a mesterséges intelligenciával való visszaélések

[8] https://academicintegrity.org/aws/ICAI/pt/sp/values

száma rohamosan nő az akadémiai közegben (Tatzel és Mael, 2023). Tovább fokozza a problémát az a tény, hogy onnantól kezdve, hogy egy szerző mesterséges intelligencia által generált szöveget használ publikációjában, mások hitelességét, szakmai integritását is veszélybe sodorhatja, hiszen az esetleges társszerzők, a szerkesztők és szakmai elbírálást végző személyek is nevüket adják a munkához (Miao, 2023).

Plágium vagy sem?

Egyes hangok szerint a nagy nyelvi modellek által generált szövegek használata etikai szürke zónaként értelmezhető, ugyanis nem állapítható meg egyértelműen, hogy hol ér véget az ember munkája, és honnantól értékelünk valamit plágiumként (Meyer és munkatársai, 2023). Hutson (2024) szintén hangsúlyozza, hogy ezen határvonalak definiálása problematikus lehet. Gondoljunk csak bele milyen sokféleképpen kooperálhat egymással ember és gép. Egy egyetlen prompt [9] alapján generált, és egy teljes mértékben sajátkezűleg írt, de végső ellenőrzésként ChatGPT-vel átolvastatott szöveg is beletartozhat ebbe a kategóriába, így valóban kérdésként merülhet fel, hogy milyen mértékű mesterséges intelligencia szerepvállalást értelmezünk plágiumként. Mindenekelőtt érdemes lehet tisztázni a plágium fogalmát, aminek bár pontos jogszabályi definíciója országonként eltérő lehet, általános jelentése a következő: más szerző munkájának felhasználása annak megfelelő hivatkozása nélkül, azaz más szerző szellemi termékének eltulajdonítása. Ha egy kicsit szabadabb gondolkodással közelítjük a témát és hátra lépünk egyet, azon a kérdésen is érdemes lehet elgondolkodnunk, hogy plágiumként értelmezhető-e egyáltalán a mesterséges intelligencia bármilyen mértékű bevonása a szövegalkotási folyamatba. A témával foglalkozó szerzők körében megoszlanak a vélemények az utóbbi kérdést illetően. Jarrah és munkatársai (2024) ide kapcsolódó szakirodalmi áttekintésének fő konklúziója az, hogy ez leginkább értelmezés kérdése. Egyes szerzők a nagy nyelvi modellek nyújtotta lehetőségek pozitívumait emelik ki, és hangsúlyozzák, hogy használatukkal emelkedett minőségű tudományos tartalom állítható elő, a szövegalkotást támogató eszközökként tekintenek rájuk (Jarrah és munkatársai, 2024). Mások szerint csak akkor vonható be mesterséges intelligencia az írás folyamatába etikus módon és a plágium vétségének felmerülése nélkül, ha precízen és alaposan jelölve van, hogy pontosan milyen módon járult az hozzá a szöveg létrejöttéhez (a szöveg aktuális részeinél

[9] a generáláshoz használt intstrukció szövege

forrásként van megjelölve), és minden ilyen módon generált információ valós forrásokkal van alátámasztva (melyek szintén hivatkozva vannak) (Jarrah és munkatársai, 2024). Meyer és munkatársai (2023) szerint ha szigorú szemlélettel közelítjük a kérdést, akkor a mesterséges intelligencia által generált tartalom felhasználása kimeríti a plágium fogalmát, ugyanis a modellek az interneten már létező információkat felhasználva konstruálják válaszaikat, azonban hogyha a generált szöveg többszörös iteráció eredménye, mely során a felhasználó folyamatosan finomítja, specifikálja prompt-jait, akkor az így megszületett szöveg már a felhasználó saját munkájaként értelmezhető és nem merül fel plágium vétsége.

Irányelvek, szabályozások

Ami biztos, hogy a plágium kérdésköre a mesterséges intelligencia fejlődésével egyre inkább előtérbe kerül, és az akadémiai közeg előtt álló fontos feladat annak újbóli definiálása az AI generált tartalom kontextusában (Hutson, 2024). Ennek okán számos oktatási intézmény, vagy tudományos munkával kapcsolatos szervezet, kiadó alkotta meg saját a témához kapcsolódó irányelveit az elmúlt években. Nem is egészen pontos itt a múlt idő használata, hiszen a terület gyors fejlődése miatt az irányelvek is folyamatos aktualizálásra szorulnak. Az egyik legjelentősebb kérdés a mesterséges intelligencia által generált tartalmak felhasználását illetően az, hogy milyen módon szükséges azt jelölni a publikációban. Lehet-e szerző a mesterséges intelligencia, hivatkozható-e szerzőként úgy, mintha egy másik tanulmány szerzőjére hivatkoznánk, vagy eszközként érdemes róla említést tenni? Ahhoz, hogy ezekre a kérdésekre választ kapjunk, először is érdemes lehet megvizsgálni a szerzőség kritériumait. Az ICMJE (International Committee of Medical Journal Editors) előírása[10] szerint ehhez négy feltételnek kell teljesülnie: jelentős hozzájárulást kell tenni a munka valamely részéhez; részt kell venni a tanulmány szövegének megalkotásában, vagy annak kritikai felülvizsgálatában; jóvá kell hagyni a végleges munka publikálását és felelősséget kell vállalni a munka minden aspektusáért (itt érdemes említést tenni róla, hogy ezeket a kritériumokat más szervezetek, kiadók is átvették/alkalmazzák, például a Springer). Ezek alapján az a következtetés vonható le, hogy problémás lehet nagy nyelvi modelleket szerzőként megjelölni, hiszen bár jelentős hozzájárulást képesek lehetnek tenni a publikációs munkához, de jóváhagyni azt, vagy felelősséget vállalni érte már nem áll módjukban (Ide és munkatársai, 2023). Ezzel megegyező

[10] https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

a legtöbb tudományos szaklap álláspontja, konszenzus látszódik tehát azt illetően, hogy mesterséges intelligencia alapú program nem fogadható el tudományos publikáció szerzőjeként, és ebből fogva nem is hivatkozható szerzőként a felhasznált irodalmak között. Kevésbé mutatkozik egyetértés az AI által generált tartalom felhasználását illetően. Az Elsevier kiadó irányelvei[11] például megengedik a jobb olvashatóság és szebb fogalmazás érdekében történő mesterséges intelligencia használatot, de azt is specifikálják, hogy ennek a szerző alapos felülvizsgálatával kell történnie. Ezzel megegyező a Springer előírása[12] is, miszerint a szöveg stílusának és olvashatóságának javítása érdekében szerkesztési céllal használható mesterséges intelligencia, és ebben az esetben nem szükséges annak megjelölése. Hangsúlyozzák azonban, hogy az AI ezen módon történő hasznosításának nem lehet része tartalomgenerálás. Az Amerikai Pszichológiai Társaság (APA) közleménye [13] az eddigiekkel többnyire egybehangzó, ők sem ismernek el szerzőként semmilyen nagy nyelvi modellt, valamint előírják, hogy használatukat speciális formában szükséges hivatkozni („software citation template"), amely tartalmazza, hogy a szerző milyen módon és milyen céllal használt mesterséges intelligenciát (előírás továbbá az output-ok csatolása mellékletként). Bár többnyire nem szigorú előírásként jelenik meg, de ezek a közlemények gyakran kitérnek arra is, hogy nem javasolt a tanulmányok teljes szövegét input-ként megadni nagy nyelvi modelleknek, fokozottan igaz ez szakmai elbírálás, társértékelés esetén. A tudományos szaklapok, kiadók mellett legtöbbször az oktatási intézmények is rendelkeznek AI használatra vonatkozó iránymutatással. Az ELTE esetében a Pedagógiai és Pszichológia kar fogalmazott meg részletes útmutatót[14], amely 2023 áprilisa óta van érvényben, de 2024-ben bővítették azt a szerzők.

I.4 A detekció szükségessége

A mesterséges intelligencia fejlődésével és a különböző nagy nyelvi modellek megjelenésével az előzőekben bemutatott módon lehetőségek mellett aggályok is felmerülnek. Ahogy azt Sikander és munkatársai (2023) eredményei is mutatják, bizonyos LLM-ek szövegalkotási készsége mára olyan szintre emelkedett, hogy nehéz feladatnak bizonyul megkülönböztetni az

[11] https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals
[12] https://www.springernature.com/gp/policies/editorial-policies
[13] https://www.apa.org/pubs/journals/resources/publishing-tips/policy-generative-ai
[14] https://www.ppk.elte.hu/dokumentumok/mesterseges-intelligencia

ilyen módon generált tartalmat az ember írta szövegtől. A különböző tudományos kiadók, szaklapok és oktatási intézmények által megfogalmazott mesterséges intelligencia használatára vonatkozó előírások miatt a detekció témája is egyre inkább előtérbe került. Chakraborty és munkatársai szerint (2023) az ilyen detektorok a nagy nyelvi modellek etikátlan felhasználása ellen „védőeszközként" szolgálhatnak.

Wu és társai (2025) öt szempontból közelítik meg azt, hogy miért is fontos és szükséges feladat a mesterséges intelligencia által generált szövegek detekciójával foglalkozni. A nagy nyelvi modellek működése és használata számos jogi kérdést vet fel elsősorban a szellemi tulajdont illetően, ezért a szabályozások betartatása és a potenciális visszaélések visszaszorítása miatt felértékelődhet a detektorok szerepe. Egy másik probléma, hogy az interneten egyre nagyobb mértékben kéretlenül is a felhasználókra áradó AI tartalom egyfajta bizalomvesztéshez vezethet, ami szintén rávilágít az ilyen jellegű tartalmak azonosításának fontosságára. Az akadémia integritás megőrzésében is kulcsszerepet játszhatnak detektorok, hiszen az etikátlan AI használat roncsolhatja a tudományos munkával kapcsolatban kialakított képet a köztudatban. Rendkívül érdekes gondolat, hogy az AI tartalmak elterjedése magukra a tartalmakat generáló modellekre is negatív hatással lehetnek hosszabb távon, ha azok későbbi, továbbfejlesztett verziói már az elődeik által előállított információn is tanulnak, ezért a mesterséges intelligencia fejlesztéssel foglalkozó vállalatok érdeke is, hogy modelljeik tanuló adatbázisából képesek legyen kiszűrni a nem embertől származó tartalmakat. Esetlegesen társadalmi szinten érezhető következményei is lehetnek a mesterséges által generált szövegek térnyerésének, azok ugyanis romboló hatással lehetnek kommunikációs és nyelvi sokszínűségünkre (Wu és munkatársai, 2025)

I.5 Eddigi eredmények klasszikus gépi tanulási algoritmusokkal

Szakdolgozatomban azt vizsgálom, hogy hagyományos klasszifikációs algoritmusok milyen hatékonysággal képesek mesterséges intelligencia által generált szöveget detektálni. A következőben néhány hasonló céllal végzett kutatás eredményeit mutatom be, melyek kiindulási alapot adhatnak saját kutatásomhoz.

Canhasi és Shijaku (2023) mesterséges intelligencia által generált szöveg detektálásról szóló tanulmányukban TOEFL nyelvvizsga esszéket vettek alapul. Minden valódi, ember által készített

esszéhez ChatGPT segítségével generáltak mesterséges verziót az eredeti kérdések alapján, tehát szövegpárok használtak. A szövegekből két különböző módszerrel vontak ki numerikus változókat. A TF-IDF vektorizácó mellett különböző karakter, szó és mondat szintű, valamint POS-hoz [15] kapcsolódó statisztikai mutatókat alkalmaztak. A szerzők által választott klasszifikációs algoritmus az XGBoost volt, mellyel mindkét változócsoportot felhasználva készítettek egy-egy modellt. Keresztvalidációs kiértékelésük eredményeként azt kapták, hogy mindkét modell kiemelkedő pontossággal prediktál, a TF-IDF változókon tanított 98%-os, míg a statisztikai mutatókat alkalmazó XGBoost modell 96%-os „accuracy" értéket ért el. A legnagyobb magyarázó erővel bíró változónak az egyedi szavak száma bizonyult, mely alapján minél több különböző szó szerepelt egy szövegben, annál valószínűbb volt, hogy a szöveget a ChatGPT generálta (Canhasi és Shijaku, 2023).

Rashid és Sodhi (2024) hasonló kutatásukban szintén mesterséges intelligencia generált és ember által írt szövegek klasszfikációjával foglalkoztak. Ők is TF-IDF súlyozású gyakoriságértékeket vontak ki a nyers szövegekből, majd pedig egy Naive Bayes és egy logisztikus regressziós modellt tanítottak az így kapott változókat alapul véve. Azt találták, hogy a két algoritmus közül 98 százalékos prediktálási pontossággal a logisztikus regresszió alkalmasabb AI generált szöveg detektálására (Rashid és Sodhi, 2024).

Nguyen és társai (2023) megközelítése az eddigiektől valamelyest eltér, ugyanis ők a klasszifikációs algoritmusok hatékonyságán kívül az ember által írt és az mesterséges intelligencia által generált szövegpárok közötti hasonlóságot is vizsgálták. Két eltérő területről származó szövegeket vettek alapul kutatásukhoz, újságcikkeket és Wikipédia oldalak tartalmát elemezték. Több különböző módon vontak ki numerikus változókat a szövegekből (TF-IDF, Bag of Words, statisztikai mutatók), majd pedig főkomponenselemzés segítségével redukálták a dimenziók számát. Az így kapott végleges változókkal Random Forest, SVM és XGBoost modelleket tanítottak. Eredményül azt kapták, hogy a Random Forest és az XGBoost modellek 99% feletti „accuracy" értékekkel szinte tökéletes predikcióra képesek, míg az SVM modell valamelyest gyengébben teljesít (74%). A változók utólagos kiértékelése alapján jelentős szerepe lehet a szósűrűségnek, a szövegek Coleman-Liau értékének, a szószámnak, de a

[15] part of speech

nyelvtani és betűzési hibák mennyiségének és a nagybetűvel kezdőd szavak gyakoriságának is (Nguyen és munkatársai, 2023).

Shah és szerzőtársai (2023) szintén Wikipédia cikkeket használtak kutatásukhoz. A szövegek AI verziót két különböző nagy nyelvi modellel is elkészítették (GPT J, Orca), így nem csak az tudták vizsgálni, hogy modelljeik milyen pontossággal tudják felismerni, hogy milyen szerző készítette a szöveget, hanem azt is, hogy melyik nagy nyelvi modell esetében hatékonyabb a detekció. Numerikus változókként lexikai jellemzőket, olvashatósági értékeket (Flesch Reading Ease, Flesch-Kincaid Grade level, Gunning Fog Index, stb.) valamint a szóhasználat gazdagságára és változatosságára vonatkozó értékeket (Yule's Characteristic K, Herdan's C, stb.) vettek alapul. Az így kapott változókkal logisztikus regresszió, döntési fa, SVM, Random Forest és XGBoost modelleket tanítottak. Eredményeik azt mutatják, hogy az Orca nagy nyelvi modellel generált szövegek esetén összességében pontosabb volt a predikció, a logisztikus regresszió itt 92%-os míg az SVM 91%-os accuracy értéket ért el. A GPT-J-vel generált adatok esetén a Random Forest teljesített a legjobban 78%-os predikciós pontossággal. Az eredmények interpretálhatósága céljából végzett LIME és SHAP elemzésekből kiderül, hogy a két legfontosabb változónak a Herdan's C érték és a Simpson index bizonyultak. Előbbi azt méri, mennyire sokszínű a szöveg szóhasználata, utó pedig azt mutatja meg, hogy mennyire jellemző a szövegre a szavak ismétlődése. Minél magasabb értéket vesz el ezeken a változókon az adott szöveg, annál valószínűbb, hogy a szerző egy nagy nyelvi modell. Ez úgy értelmezhető, hogy bár a mesterséges intelligencia által generált tartalom érezhetően gazdag szókincscsel rendelkezik, ezen a gazdag szókincsen belül gyakori az ismétlődés (Shah és munkatársai, 2023).

Durak és munkatársai (2023) szintén több nagy nyelvi modellt vizsgáltak a detekció hatékonysága szempontjából. Diákok esszéit vették alapul, majd az esszékérdéseket prompt-ként felhasználva szintetikus szövegeket generáltak három LLM-et felhasználva (BingAI, ChatGPT és Gemini). Különböző statisztikai mérőszámokat kalkuláltak az esszékhez, majd ezek alapján összahasonlították az azonos szerzőhöz tartozó szövegeket (itt az emberre, mint egy szerzőre gondolok). Ezt követően TF-IDF algoritmussal vektorizálták a szöveget. Kifejezetten fa-alapú együttes tanulási algoritmusok teljesítményét vizsgálták, ezért a kivont TF-IDF értékekkel következő modelleket tanították: Random Forest, AdaBoost, XGBoost, Bagging, Extra Trees. Összességében azt találták, hogy a Gemini által generált szövegek esetében a legpontosabb a detekció (minden modell esetében 90% fölötti accuracy, míg a ChatGPT esetében valamivel

szerényebb teljesítményre képesek az említett algoritmusok (minden modell esetében 80% és 90% közötti pontosság).

Egy másik kutatásban André és munkatársai (2023) tudományos cikkek absztraktjait használták fel, azzal a kitétellel, hogy matematikai képleteket tartalmazó és/vagy 500 karakternél alacsonyabb hosszúságú absztraktokat nem vettek figyelembe. Ezekhez a ChatGPT 3.5 Turbo nevű nagy nyelvi modellel generáltak szintetikus párokat, így állt össze a végleges korpusz. Két megközelítéssel vontak ki változókat a szövegből. Egyfelől olyan statisztikai, stilisztikai és lingvisztikai mutatókat alkalmaztak, mint a perplexity, a Type-Token arány, az átlagos token hosszúság vagy a nyelvtani hibák száma a tokenek számával arányosítva, másfelől pedig TF-IDF értékeket használtak változókként. Klasszifikációs algoritmusként logisztikus regressziót, Random Forest-t és Multinomial Naive Bayes-t használtak. A statisztikai mutatókon alapuló modellek esetében azt találták, hogy mindhárom algoritmus magas pontossággal prediktál (>97%), de a leghatékonyabb detektornak a Random Forest bizonyult (98,4%). A TF-IDF változók esetében 98,8%-os „precision" értékével a logisztikus regresszió volt a leghatékonyabb. A legpontosabb statisztikai mutatókon alapuló modellt (Random Forest) további elemzésnek vetették alá a szerzők azt vizsgálva, hogy mely változóknak jut jelentős szerep a predikciós döntés meghozatalában. Ez alapján elmondható, hogy kiemelkedő szerepe van a perplexity mutatónak, amely azt méri, hogy egy nagy nyelvi modell milyen pontossággal képes megjósolni egy szöveg egymás után következő szavait, egyszerűbben tehát azt, hogy mennyire kiszámítható egy szöveg. Ennek kiszámítására André és munkatársai (2023) az OpenAI GPT-2 modelljét használták és azt találták, hogy minél alacsonyabb egy szöveg perplexity értéke, azaz minél kiszámíthatóbb, annál valószínűbb, hogy a szerző a mesterséges intelligencia volt. Alacsonyabb mértékben ugyan, de a nyelvtani hibák is jelentőséggel bírnak egy szöveg szerzőjének megítélésében (AI vagy ember) (André és munkatársai, 2023).

Az imént bemutatott tanulmányok alapján elmondható, hogy hatékonyan detektálható mesterséges intelligencia által generált szöveg tradicionális gépi tanulási algoritmusokkal, melyek gyakran 95% fölötti predikciós pontossági értékeket is elérnek. A klasszikus szövegvektorizációs algoritmusok (TF-IDF, n-gram gyakoriság) mellett megfontolandó egyéb megfigyelt statisztikai mutatók használata is, ezek ugyanis több esetben is a detekció döntő faktorainak bizonyultak. Ilyen nagy jelentőséggel bíró mutatók lehetnek többek között különböző kiszámíthatósági, olvashatósági vagy a szókincs gazdagságára utaló mérőszámok, de

a nyelvtani hibák számának is fontos szerepet lehet. Az ideális klasszifikációs algoritmust illetően nincs teljes konszenzus a szakirodalomban, de az látható, hogy a logisztikus regresszió és az XGBoost több esetben is hatékony detektorként tűnt fel. Fontos azonban kiemelni, hogy minden itt bemutatott kutatásban a szerzők valamilyen specifikus területről származó szövegállományt (újságcikkek, Wikipédia oldalak szövegei, kutatási absztraktok) használtak modelljeik tanítására és tesztelésére, a magas predikciós pontosságokat ennek fényében kell értékelni. Más szóval ezen eredményekből nem következik az, hogy a klasszikus gépi tanulási algoritmusok általánosan, területtől függetlenül bármilyen típusú szöveg esetében hatékonyan tudnának detektálni, továbbá az eredményekből az is világosan látszik, hogy lehetősége lehet annak is, hogy melyik nagy nyelvi modell generálta a vizsgálni kívánt szöveget.

II. Módszertan

II.1. Áttekintés

Dolgozatomban azt vizsgálom, hogy hagyományos gépi tanulási klasszifikációs algoritmusok milyen hatékonysággal képesek mesterséges intelligencia által generált szöveget detektálni a modern transzformer alapú detektorokhoz képest. A kutatás során tudományos publikációk absztraktjait, és azok mesterséges intelligencia által generált verzióit veszem alapul. A nyers szövegekből a tisztítás és az előfeldolgozás után több módon vonok numerikus változókat. Egyfelől a korpusz uni- és bigramjait felhasználva TF-IDF és Count vektorizációt alkalmazok, másfelől pedig különböző olvashatóságra, repetitivitásra és szóhasználatra vonatkozó mérőszámokat kalkulálok. A vektorizációval kapott változókkal logisztikus regressziós, Naive Bayes és XGBoost modelleket tanítok, az egyéb módon kinyert változók mentén pedig összehasonlítom a mesterséges intelligencia által generált, és az emberi szerzők által írt szövegeket. A tanított modellekkel és egy nem általam tanított transzformer alapú detektorral predikciót végzek a teszthalmazon, ezt követően pedig pontosság, f1, precízió és visszahívás értékek alapján vetem össze a kapott eredményeket. Végezetül SHAP elemzéssel megvizsgálom, hogy a legjobban teljesítő modelljeim esetében milyen változók játszanak fontos szerepet a predikciós döntés meghozatalában. Kitekintésként azzal is foglalkozok, hogy egy általam készített kis elemszámú mintán más nagy nyelvi modellek által generált absztraktokon is letesztelem a ChatGPT szövegein tanított modelljeimet.

II.2. Felhasznált adat

Mivel kutatásom tágabb kontextusa a mesterséges intelligencia szerepének tanulmányozása az akadémia világában, a kutatási kérdésemben megfogalmazott detekciós feladatot specifikusan akadémiai szövegeket felhasználva végzem el.

Az adatbázis, melyet szakdolgozatom során használok, egy interneten elérhető, korábbi egyetemi szakdolgozathoz készült adatbázis [16] . 10 000 kutatási absztraktot, és azok mesterséges intelligenciával generált verzióit tartalmazza, tehát összesen 20 000 sor található benne. Az eredeti absztraktok egy korábbi, ingyenesen hozzáférhető kifejezetten kutatási absztraktokat tartalmazó adatbázisból [17] származnak (arxiv-abstracts-2021) (Clement és munkatársai, 2021). A kutatásom során használt adatbázis szerzői az eredeti, közel 2 millió absztraktot tartalmazó adatbázisból olyan módon vettek mintát, hogy a mintában a szövegek szószáma egyenletes eloszlást kövessen. Az absztraktok mesterséges intelligencia által generált verzió a ChatGPT segítségével készültek el, a generáláshoz használt pontos verzió a GPT-3.5-turbo-0301 volt (2023.03.01, 175 milliárd paraméter). A generálás az eredeti absztrakt címe és szószáma alapján történt. A folyamat automatizálása érdekében a szerzők a modell API verzióját használták. Az API dokumentációja szerint a GPT-3.5-turbo-0301 nem tulajdonít túl nagy jelentőséget a kérésbe írandó rendszerüzenetnek, ezért a szerzők felhasználói üzenet részben specifikálták a generáláshoz elsődlegesen használni kívánt prompt-ot (Sivesind és Winje, 2023). Az eredetileg felhasznált API „request” az alábbi ábrán látható:

[16] https://huggingface.co/datasets/NicolaiSivesind/human-vs-machine?utm_source=chatgpt.com
[17] https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021

```
"messages": [
    {"role": "system",
     "content":
        "You are a helpful assistant which produces
         ↪ abstracts for research papers based on a title
         ↪ and a length. Your task is to produce the
         ↪ abstract which suits the title, and is of the
         ↪ desired length in words."},

    {"role": "user",
     "content":
"""Title: \"{title}\"
Abstract length: {word_count_goal} words

---

Above is the title of a scientific research paper and the
 ↪ length of its abstract. Using a formal/academic
 ↪ language, write a new abstract which matches title and
 ↪ has a length of {word_count_goal} words. Do not answer
 ↪ with anything else than the abstract."""}]
```

1. Ábra - A szintetikus absztraktok generálásához felhasznált prompt-ot tartalmazó API „request” (Sivesind és Winje, 2023)

II.3. Tisztítás, előfeldolgozás

Az előfeldolgozás teljes folyamatát Python[18] programozási nyelv segítségével végeztem el, a használt modulokat és függvényeket az egyes lépéseknél ismertetem.

Az adatok áttekintése során észrevettem, hogy számos absztrakt matematikai képleteket is tartalmaz. Ezt követően megvizsgáltam, hogy az ember által írt, és az AI generált szövegeknél milyen mértékben vannak jelek képleteket tartalmazó absztraktok, és azt találtam, hogy sokkal jellemzőbb ez a jelenség az emberi szövegek esetén. Lehetséges megoldásként felmerült, hogy a képleteket egy „placeholder” tokennel helyettesítem, de arra a döntésre jutottam, hogy nem szeretnék a képleteknek túl nagy jelentőséget adni, ezért azokat az absztrakt párokat, ahol legalább az egyik szöveg matematikai formulát tartalmazott, eltávolítottam az adatbázisból. Ezeket azért tudtam egyszerűen beazonosítani, mert a LaTeX-ben íródott képletek mindig „$” jelek között szerepelnek a szövegekben.

[18] a szakdolgozat során felhasznált kódok az alábbi linken érhetők el: https://github.com/maklaryj/thesis_ma

Tekintettel arra, hogy kutatási absztraktok és azok mesterséges intelligencia által generált verzió alkották a kutatásom során használt korpuszt, nem volt szükség hosszadalmas és komplex tisztítási fázisra. Először eltávolítottam minden olyan nem alfanumerikus karaktert, amely nem alfanumerikus vagy „whitespace" volt. Ezt követően kiszűrtem az új sort jelző karakterpárokat, és a felesleges „space" karaktereket is töröltem minden szövegből. Ezeket reguláris kifejezés alapú minták segítségével valósítottam meg, melyhez a re [19] modult használtam.

A tisztítási fázis után tokenizáltam, azaz kisebb egységekre bontottam a szövegeket. Ezt a műveletet több szinten el lehet végezni, kutatásom esetén a szó szintű tokenizáció volt a legindokoltabb. Ennek megvalósításához az nltk modul word_tokenize [20] függvényét használtam, amely egy string-eket tartalmazó listává alakítja szövegeket.

Ezt követően egy lépésben valósítottam meg a lemmatizálást és a stop szavak eltávolítását. A lemmatizálás (vagy szótövezés) a természetesnyelv feldolgozás egy olyan alapvető művelete, mely során a szó ragozott, toldalékolt formáját visszavezetjük annak szótári alakjára (Dave és Balani, 2015). Ennek eredményeképp a hasonló szótővel rendelkező szavak egyenértékű tokenekként kerülnek feldolgozásra. A szótövezés minősége nagyban múlik a megfelelő POS (Part-of-Speech) címkék használatán, azaz a tokeneket szükséges ellátni azzal az információval, hogy milyen szófajba tartoznak (Karwatowski és Pietron, 2022). Kutatásomban a szótövezést az nltk modul WordNetLemmatizer[21] függvényével valósítottam meg, ami alapvetően nem végez POS címkézést, így azt a szótövezést megelőzően külön végeztem el az nltk modul pos_tag[22] függvényének segítségével.

Stop szavak alatt a gyakran előforduló, de önmagukban kevés információs tartalommal rendelkező szavakat értjük, melyek eltávolításával a lényegi információ jut jelentősebb szerephez a szövegben, és amennyiben vektorizáljuk a korpuszt, jelentősen csökken a vektorizáció dimenziónak száma (Rani és Lobiyal, 2022). Ilyen szavak lehetnek például a névelők, névmások, kötőszavak vagy segédigék. Kutatásomban az nltk modul angol nyelvű stopwords listáját alapul véve távolítottam el stop szavakat a korpuszból.

[19] https://docs.python.org/3/library/re.html
[20] https://www.nltk.org/api/nltk.tokenize.word_tokenize.html
[21] https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=wordnet
[22] https://www.nltk.org/api/nltk.tag.pos_tag.html

II.4 Változókivonás

II.4.1 Vektorizáció

A szöveganalatikai kutatások és elemzések elengedhetetlen része a nyers, struktúrálatlan szövegek struktúrált adattá alakítása. Ezt a folyamatot, mely során a szöveget számokat tartalmazó vektorokká alakítjuk, vektorizációnak nevezzük. A vektorizációra azért van szükség, mert a szöveganalitikában használatos algoritmusok jelentős része numerikus változókkal dolgozik. Létezik a vektorizációnak olyan fajtája is, amikor a nyers szöveget nem rögtön számokat, hanem először string formátumú elemeket tartalmazó vektorokká alakítjuk, de szakdolgozatomban kizárólag direkt numerikus vektorizációs eljárásokat alkalmazok, szám szerint kettőt (Bag of Words, súlyozott Bag of Words: TF-IDF). Fontos továbbá megjegyezni, hogy vektorizáció előtt tokenizálni kell a szöveget, de ezt bizonyos függvények integráltan megteszik (Krzeszewska és munkatársai, 2022).

Bag of Words

A Bag of Words eljárás a szövegvektorizáció legrégebbi és egyik leggyakrabban alkalmazott módja. Ez a modell a szöveget úgy értelmezi, mint szavak összességét, bármiféle logikai vagy nyelvtani kapcsolatot figyelmen kívül hagyva, a neve is innen ered, miszerint a szövegre, mint egy zsák szóra tekinthetünk (Abubakar és munkatársai, 2022). Az eljárás lényege, hogy a korpuszban található egyedi tokenek számával megegyező hosszúságú vektorral reprezentálja az egyes dokumentumokat, mely vektor elemeit az egyedi tokenek gyakoriságai adják. A Bag of Words vektorizáció alapvetően szavakat, azaz unigramokat vesz alapul, de létezik kevésbé szigorú megközelítés is, mely szerint bármilyen n-grammal alkalmazható az eljárás (Bag of N-grams, n-gram alatt egy n hosszúságú tokenláncot értünk).

Szakdolgozatomban a Python programozási nyelv scikit-learn moduljának CountVectorizer[23] függvényét használtam a Bag of Words vektorizáció megvalósítására. Az n_gram_range paraméter különböző értékekre történő beállításával több módon is vektorizáltam a szövegeket.

TF-IDF

[23] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

A TF-IDF egy széles körben alkalmazott súlyozási módszer, amely alapvetően a Bag of Words vektorizációra épül, de figyelembe veszi azt is, hogy az egyes tokenek mennyire gyakoriak a teljes korpuszban. Ezáltal csökken az általánosan gyakori szavak relevanciája, és jobban előtérbe kerülnek az egyes dokumentumokban meghatározó szerepet játszó informatív tokenek, amelyek a teljes korpuszban ritkábban fordulnak elő (Manning és munkatársai, 2008). A súlyozott gyakoriságértékek kiszámítása a következőképpen zajlik:

*TFIDF(szó, dokumentum) = TF(szó, dokumentum) * IDF(szó)*

Ahol:

- *TF(szó, dokumentum*) = szó gyakorisága a dokumentumban / összes szó száma a dokumentumban
- *IDF(szó)* = log(1 + (dokumentumok száma/ dokumentumok száma, amelyek tartalmazzák a szót))

(Das és Chakraborty, 2018)

A TF-IDF súlyozott gyakoriságok kiszámításához szintén a Pythonhoz készült scikit-learn modult használtam, azon belül pedig a TfidfVectorizer[24] függvényt. Ahogy a Bag of Words modell esetében, az n_gram_range paraméter finomításával itt is többféle vektorizációt végeztem.

II.5. Klasszifikációs algoritmusok

A szakirodalomban fellelhető korábbi eredmények alapján több különböző klasszifikációs algoritmust is alkalmazok vizsgálatom során. Számos korábbi hasonló témájú kutatásban is magas predikciós pontosságot értek el logisztikus regressziós és XGBoost modellekkel, ennek okán ezeket én is bevonom elemzésembe. Valamint annak érdekében, hogy működési elvük tekintetében változatos legyen az alkalmazott osztályozók csoportja, Naive Bayes modelleket is tanítok.

A logisztikus regresszió megvalósításához az scikit-learn csomag LogisticRegression [25] függvényét használom, a Naive Bayes esetén ugyanezen modul MultinomialNB[26] függvényét

[24] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[25] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[26] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

veszem alapul, az XGBoost modellek létrehozását az xgboost csomag XGBClassifier [27] metódusával valósítom meg.

II.5.1 Logisztikus regresszió

A logisztikus regresszió egy statisztikai alapokon nyugvó felügyelt tanulási módszer. Elsősorban bináris klasszifikációs problémáknál alkalmazható, azaz olyan esetekben, amikor a prediktálni kívánt célváltozónak két lehetséges értéke van. A módszer nem sorolja közvetlenül a két osztály valamelyikbe az aktuális megfigyelést, azaz nem hoz determinisztikus döntést, hanem az egyes osztályokba tartozás valószínűségét modellezi, tehát a logisztikus regressziós modell kimeneti értéke mindig a $[0,1]$ intervallumba esik (James és munkatársai, 2013).

Matematikai háttér

Ha a valószínűséget szeretnénk a prediktor változók valamilyen lineáris kombinációjaként felírni, azzal a problémával szembesülünk, hogy a felvett érték bárhova eshet a $[-\infty, \infty]$ tartományon belül. A logisztikus regresszió ennek okán nem közvetlenül a kimeneti valószínűségét modellezi, hanem annak esélyhányadosának természetes alapú logaritmusát. Ezzel a megoldással áthidalható a $[0,1]$ és a $[-\infty, \infty]$ értékkészletek közötti diszkrepancia. Az esélyhányados az esemény bekövetkezésének valószínűsége elosztva annak valószínűségével, hogy nem következik be az esemény. A log-odds a prediktor változók lineáris kombinációjaként a következőképpen írható fel:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p$$

Ahol:

- β_0 intercept (konstans), a logit érték akkor, ha minden magyarázó változó értéke 0
- β_i az i sorszámú magyarázó változóhoz tartozó regressziós együttható, amely megmutatja, hogy egy egységnyi növekedés az adott magyarázó változón hogyan változtatja meg az esélyhányados természetes alapú logaritmusának értékét
- X_i az i sorszámú magyarázó változó

[27] https://xgboost.readthedocs.io/en/stable/get_started.html

Az egyenlet átalakításával kifejezhető a prediktált valószínűség, mely szigmoid függvény alakot vesz fel. Ezzel a transzformációval biztossá válik, hogy a felvett értékek a $[0,1]$ intervallumban maradnak. Az említett szigmoid függvény a következőképpen írható fel:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p}}$$

Az együtthatók becslésére alapvetően több lehetőség is rendelkezésre áll, de logisztikus regresszió esetén a maximum-likelihood módszer a leginkább alkalmazott. Ennek lényege, hogy a rendelkezésre álló adatokat felhasználva olyan regressziós együtthatókat találjunk, melyekkel a becsült valószínűségek a lehető legközelebb kerülnek a magyarázott változón felvett tényleges értékekhez. A logisztikus regresszió együtthatóbecslésének likelihood függvénye egyváltozós esetben a következőképpen írható fel:

$$\ell(\beta_0, \beta_1) = \prod_{i:\, y_i = 1} p(x_i) \prod_{i':\, y_{i'} = 0} \left(1 - p(x_{i'})\right)$$

A predikció során a becsült valószínűséget egy előre meghatározott küszöbértekhez hasonlítjuk. A becsült valószínűséget a fent említett $p(X)$ -re kifejezett egyenlet segítségével kapjuk meg. A predikció 0,5-ös küszöbértéket használva a következőképpen írható fel:

$$\hat{y} = \begin{cases} 1 \text{ ha } \hat{p}(X) \geq 0{,}5 \\ 0 \text{ ha } \hat{p}(X) < 0{,}5 \end{cases}$$

Ez azt jelenti, hogy amennyiben a becsült valószínűség nagyobb vagy egyenlő, mint a meghatározott 0,5-ös küszöbérték, akkor a megfigyelés prediktált értéke a kimeneti változón 1-lesz, ellenkező esetben pedig 0.

(James és munkatársai, 2013).

II.5.2 Naive Bayes

A Naive Bayes egy szöveganalitikában gyakran alkalmazott egyszerű, de hatékony klasszifikációs algoritmus, amely bináris és több osztályos helyzeteken egyaránt működik. A

módszer Bayes tételét veszi alapul, azzal az erős feltételezéssel, hogy a prediktor változók függetlenek egymástól az egyes osztályokon belül. A módszer neve onnan ered, hogy ez a feltételezés gyakran nem teljesül. Amellett, hogy sok esetben magas predikciós pontosság érhető el Naive Bayes-t használva, számítási szempontból is hatékonynak mondható, melyek okán széles körben elterjedt az algoritmus használata (Webb és munkatársai, 2011).

Matematikai háttér

A Naive Bayes algoritmus tehát Bayes tételét felhasználva határozza meg azt, hogy egy adott megfigyelés adott prediktor változókon felvett értékek mellett melyik osztályba tartozik legnagyobb valószínűséggel. A tétel egy adott k osztály esetén a következőképpen írható fel:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k) \cdot P(Y = k)}{P(X = x)}$$

Ahol:

- $P(Y = k|X = x)$ a poszteriori valószínűség, azaz annak valószínűsége, hogy az az adott prediktorváltozó értékek mellett a megfigyelés a k osztályba tartozik
- $P(Y = k)$ a priori valószínűség, azaz annak valószínűsége, hogy a megfigyelés a tanuló adatok alapján a k osztályhoz tartozik
- $P(X = x|Y = k)$ a likelihood, ami azt mutatja meg, hogyha a megfigyelés a k osztályba tartozik, akkor mennyire valószínűek a prediktor változón felvett értékek a k osztályban
- $P(X = x)$ a teljes valószínűség, azaz az összes priori valószínűség és likelihood pár szorzatainak összege

A Naive Bayes feltételezésének értelmében a prediktor változók feltételesen függetlenek, azaz az adott értékek együttes bekövetkezésének valószínűsége kiszámítható az egyenkénti valószínűségek összeszorzásával:

$$P(X = x|Y = k) = \prod_{i=1}^{n} P(X = x_i|Y = k)$$

A predikció eredményét úgy kapjuk meg, ha minden osztály esetében kiszámítjuk a poszterior valószínűségét, majd vesszük a legnagyobb poszteriori valószínűséghez tartozó osztályt:

$$\hat{y} = arg \max_{k} \left[P(X = k) \prod_{j=1}^{p} P\left(X_j = x_j \middle| Y = k\right) \right]$$

(Mitchell, 1997)

II.5.3 XGBoost

Az XGBoost egy döntési fa alapú „ensemble" algoritmus, azaz olyan gépi tanulási módszer, amely több modellt kombinál annak érdekében, hogy predikciós pontosságát maximalizálni tudja. Ellentétben a „bagging" elven működő algoritmusokkal, ahol az egyes fák egymástól függetlenül tanulnak (mint például a Random Forest), a „boosting" algoritmusok esetében sorrendben épülnek egymásra a részmodellek. Az XGBoost minden iterációban úgy épít új döntési fát, hogy az előző modellek hibáira, azaz a predikciók reziduálisaira fókuszál. Az új fák célja ezen hibák csökkentése, melynek eredményeképpen a modell minden iterációval pontosabbá válik. (Chen és Guestrin, 2016).

Az XGBoost legnagyobb előnye, hogy kiváló teljesítményre képes, melyet a modellt bemutató eredeti tanulmány szerzői azzal prezentálnak, hogy a Kaggle által meghirdetett versenyek jelentős százalékén valamilyen XGBoost-ot használó megoldás éri el az első helyet (itt fontos megjegyezni, hogy a cikk 2016-ban jelent meg). Optimalizált működésének köszönhetően alacsonyabb memóriával rendelkező környezetben is rendkívüli gyorsaságra képes, valamint, kifejezetten nagy mennyiségű adat esetén is kimagasló teljesítményt nyújt. Jól kezeli a sok hiányos cellával rendelkező úgynevezett „sparse" adatbázisokat is, és különböző beépített regularizációs technikákkal csökkenti a túlillesztés kockázatát (Chen és Guestrin, 2016).

Matematikai háttér

Az XGBoost célja minimalizálni a regularizált veszteségfüggvényt, mely a következőképpen írható fel:

$$\mathrm{L}(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k)$$

Ahol:

- $l(y_i, \hat{y}_i)$ veszteségfüggvény, amely a becsült és valódi érték közötti eltérést méri, bináris klasszifikáció esetén alapértelmezetten logaritmikus veszteség (log loss), de a feladattól függően rugalmasan választható
- $\Omega(f_t)$ a regularizációs tag, amely a modell komplexitását szabályozza

Minden iterációban új döntési fát (f_t) adunk a modellhez, amely az aktuális veszteségfüggvény alapján a legnagyobb mértékű javulást (azaz veszteségcsökkentést) eredményezi. Az egyes iterációkban a veszteségfüggvény tehát a következőképpen írható fel:

$$L(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Ahol:

- $l(y_i, \hat{y}_i)$ veszteségfüggvény, amely a becsült és valódi érték közötti eltérést méri
- $f_t(x_i)$ az aktuális iterációban tanított új döntési fa predikciója az i-edik megfigyelésre
- $\Omega(f_t)$ a regularizációs tag az aktuális fára
- t az iteráció sorszáma
- i a megfigyelés sorszáma

A regularizációs tag a következőképpen írható fel:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2$$

Ahol:

- T a levelek száma
- w_j a levelek súlyai
- γ és λ a modell komplexitását szabályozó hiperparaméterek

A célfüggvény optimalizálása érdekében az XGBoost nem közvetlenül az aktuális veszteségfüggvényt veszi alapul, hanem annak másodrendű Taylor közelítését. Az első és második deriváltakat felhasználva a következőképpen írható fel a közelítés:

$$L(t) \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Ahol:

- g_i és h_i az első és másodrendű deriváltak
- $f_t(x_i)$ a t sorszámú döntési fa predikciója az i sorszámú megfigyelésre
- $\Omega(f_t)$ regularizációs tag a t sorszámú döntési fa esetlben

Az XGBoost esetén a predikció úgy történik, hogy az egyes döntési fák kimeneti értékei összeadásra kerülnek, majd pedig az így kapott végső logit érték szigmoid függvény alkalmazásával valószínűséggé konvertálódik. A modell tehát nem csak determinisztikus döntések meghozatalára képes, hanem valószínűségi becsléseket is tud nyújtani. A predikciós formula bináris klasszifikáció esetén az alábbi módon írható fel:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Ahol:

- K az összes fa (iteráció) száma
- $f_k(x_i)$ a k sorszámú fa predikciója az i sorszámú megfigyeléshez

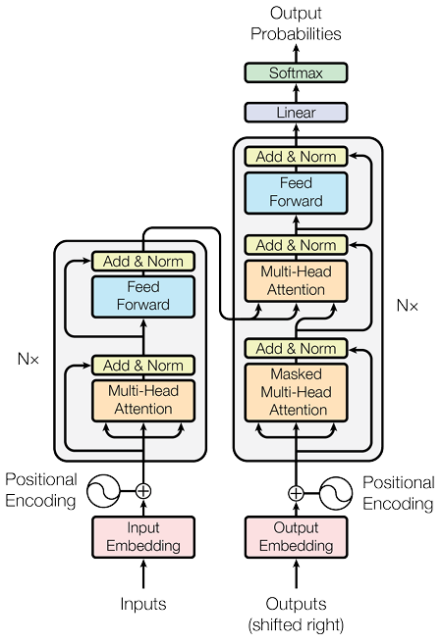
(Chen és Guestrin, 2016).

II.6 Transzformer alapú detekció

II.6.1 Transzformer architektúra

Mivel szakdolgozatomban a hagyományos gépi tanulási algoritmusok mellett egy state-of-the-art módszernek számító transzformer alapú detektort is alkalmazok, a következőkben röviden ismertetem a detektor hátterében álló architektúra felépítését.

A transzformer egy elsősorban természetesnyelv feldolgozásban alkalmazott neurális hálózati architektúra, melyet Vaswani és munkatársai mutattak be az „Attention is All You Need című tanulmányukban 2017-ben. A transzformer architektúra megjelenése a nagy nyelvi modellek szempontjából forradalmi jelentőségűnek mondható, mert a módszer a szöveg feldolgozásának olyan hatékony módját teszi lehetővé, amilyenre a korábban state-of-the-art megoldásnak számító rekurrens és konvolúciós neurális hálózatok nem képesek. A legnagyobb különbség a hagyományos neurális hálózat alapú modellek és a transzformerek között abban rejlik, hogy utóbbiak a beérkező adatot nem szekvenciálisan, hanem párhuzamosan dolgozzák fel, valamint „self-attention" mechanizmusuknak köszönhetően a szöveg minden egyes elemének kontextuális kapcsolatait figyelembe veszik. A napjainkban legnépszerűbb nagy nyelvi modellek, mint a ChatGPT, a Claude vagy a Gemini, mind transzformer architektúrára épülnek.



2. Ábra. A transzformer architektúrája (Vaswani és munkatársai, 2023)

A transzformerek architerúktája alapvetően két fő működési egységre bontható le. Az enkóder egység feldolgozza a bejövő szöveget, az így kinyert információk alapján pedig a dekóder egység generálja a modell végtermékét, az output tartalmat. Ezt a logikát használják fel például a különböző transzformer alapú chatbot-ok vagy fordítóprogramok.

Az transzformer a bemenő szöveget először tokenizálja, azaz kisebb egységekre bontja (például szavakra). Ezt követően a tokenek alapján szóbeágyazást készít, azaz minden tokenhez egy sokdimenziós vektort rendel, amely a szemantikai dimenziókon felvett értékeket tartalmazza. Mivel a transzformerek nem szekvenciálisan, hanem párhuzamosan dolgozzák fel a tokeneket, az enkóder pozícionális kódolást is végez, annak érdekében, hogy a szavak sorrendjében rejlő információt is megőrizze.

Attention

A transzformerek „lelke" az úgynevezett „self-attention" mechanizmusukban rejlik. Ez a folyamat a tokenek eredeti „embedding" vektorait olyan módon írja felül, hogy azok arról is hordozzanak információt, hogy melyik token melyik másik tokennel áll szorosabb kapcsolatban, azaz melyikre „figyel" (függetlenül attól, hogy mekkora távolságra vannak egymástól a szövegben). A „self-attention" mechanizmus tehát a tokenek kontextusba helyezett reprezentációját készíti el.

Egy szöveg „self-attention" mátrixa az alábbi képlettel számítható ki.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

A Q (query), K (key), és V (value) az eredeti tokenek „embedding" vektorai és a hozzájuk tartozó súlymátrixok (W^Q, W^K, W^V) szorzataiként kapott mátrixok, ahol a sorok száma a tokenek számával, míg az oszlopok száma az „embedding" vektorok dimenziónak számával egyezik meg. A QK^T mátrixszorzás eredményeként az egyes tokenek közötti hasonlóságértékeket kapjuk. Ezeket aztán újraskálázzuk (osztjuk$\sqrt{d_k}$-val), majd softmax függvénnyel normalizáljuk. Az így kapott értékek lesznek az úgynevezett „attention" súlyok, melyek a V (value) mátrixszal összeszorozva kiadják az egyes tokenek „attention" értékeit, azaz kontextuális reprezentációit tartalmazó „attention" mátrixot.

A transzformer modellben ezt a „self-attention" mechanizmust a „multi-head attention" blokk használja fel. A „multi-head attention" blokk lényege, hogy egyszerre több, párhuzamos „self-attention" kalkuláció megy végbe különböző súlymátrixokkal (W_i^Q, W_i^K, W_i^V). Ez azt jelenti, hogy minden ilyen „attention" fej más szempontból vizsgálja a tokenek közötti kapcsolatokat, lehetővé téve ezzel azt, hogy a modell egyszerre többféle összefüggést is felismerjen a szavak

között. Az egyes fejek kimenetei végül összefűzésre kerülnek és egy W_0 súlymátrixszal transzformálódnak annak érdekében, hogy az eredeti dimenziókba visszailleszthetők legyenek.

II.6.2 Alkalmazott transzformer alapú detektor

Hogy a kutatásom során alkalmazott klasszikus változó alapú gépi tanulási módszerek teljesítményéhez viszonyítási alapot nyújtsak, a felvázolt detekciós feladatot egy transzformer alapú módszerrel is elvégzem. A detektor[28], amit használok, egy RoBERTa Large modellre épülő kifejezetten szekvenciák (tokenláncok) bináris klasszifikációjára létrehozott modell, melyet a SuperAnnotate nevű amerikai vállalat fejlesztett. A modell tanítása során 44 ezer szövegpárt használtak (1 – embertől származó, 1 – mesterséges intelligenciától származó). Az emberi szerzőtől származó szövegeket több különböző forrásból gyűjtötték össze, így az adatbázis tartalmaz többek Wikipédia oldalakat[29], Reddit posztokat[30], és tudományos publikációkat[31] is. A szövegek AI verzióit 14 különböző nagy nyelvi modellel generálták, melyek 4 modellcsalád különböző verziói voltak (GPT, Mistral, Anthropic, Llama). A detektor teljesítménye egy kifejezetten detektorok tesztelése céljából létrehozott benchmark adatbázis [32] egy részhalmazán lett validálva mely több különböző LLM által generált szöveget is tartalmazott. Az elért átlagos predikciós pontosság 0.852 volt.

II.7 Kiértékelési metrikák

A kutatásom során használt klasszifikációs modellek és detektorok teljesítményét részben az alábbi, klasszifikációs kutatások esetén gyakran használt mérőszámokkal értékelem ki.

Pontosság

A pontosság (vagy eredeti elnevezésén „accuracy”) érték azt mutatja meg hogy az összes predikció mekkora része lett a megfelelő osztályba sorolva, bináris klasszifikáció esetén a valós pozitív és valós negatív predikciók összegének aránya az összes predikcióhoz képest (Powers, 2020). Ha az egyes osztályokban a megfigyelések száma nem azonos, a mérőszám félrevezető

[28] https://huggingface.co/SuperAnnotate/ai-detector
[29] https://huggingface.co/datasets/wikimedia/wikipedia
[30] https://huggingface.co/datasets/rexarski/eli5_category
[31] https://www.tensorflow.org/datasets/catalog/scientific_papers
[32] https://raid-bench.xyz/

lehet, más esetekben viszont egy egyszerűen értelmezhető képet ad a klasszfikáló modell teljesítményéről.

Precizitás

A precizitás (vagy eredeti elnevezésén „precision”) mérőszám az mutatja meg, hogy a pozitív predikciók mekkora aránya tartozik a valódi pozitív megfigyelések közé (Powers, 2020). Szakdolgozatom kontextusában azt számosítja, hogy az AI címkével ellátott kutatási absztraktok közül mennyi volt ténylegesen mesterséges intelligencia által generálva. Különösen fontos ez a metrika azokban az esetekben, amikor a hamis pozitív esetek minimalizálása elsődleges szempont.

Visszahívás

A visszahívás (vagy eredeti elnevezésén „recall”) azt mutatja meg, hogy a valós pozitív csoportba tartozó megfigyelések mekkora részét sorolja ténylegesen a pozitív osztályba a klasszifikáló modell (Powers, 2020), kutatásom esetében tehát azt, hogy az összes mesterséges intelligencia által generált szöveg mekkora aránya kerül detektálásra. A visszahívás mérőszámnak azokban az esetekben van különösen nagy jelentősége, ahol elsődleges fontosságú a hamis negatív esetek kiszűrése.

F1 érték

Az F1 érték egy olyan mérőszám, amely a precizitást és visszahívást egyszerre figyelembe veszi. Értékét úgy kapjuk meg, hogy vesszük az imént említett két érték harmonikus átlagát (Powers, 2020). Az F1 mutató akkor hasznos különösen, ha az egyes osztályokba tartozó megfigyelések száma jelentősen eltér egymástól. Az F1 érték képlete:

$$F1 = 2 * \frac{precizitás * visszahívás}{precizitás + visszahívás}$$

Konfúziós mátrix

A konfúziós mátrix (vagy eredeti nevén „confusion matrix”) egy klasszifikációs modellek teljesítményének kiértékelésnél használatos eszköz. Bináris klasszifikáció esetén a megfigyeléseket két dimenzió mentén osztja el két-két csoportba. Az egyik ilyen dimenzió a megfigyelés valódi osztálya, a másik pedig a prediktált osztály. Ezen felosztás eredménye a megfigyelések valós pozitív, hamis negatív, valós negatív és hamis pozitív csoportokba

rendeződése. A konfúziós mátrix nagy előnye, hogy vizuális eszközként hatékonyan szemlélteti a modell teljesítményét, és egyszerűen interpretálható. A mátrix cellái két osztály esetén a következőképpen alakulnak:

		Valós osztály	
		Pozitív	Negatív
Prediktált osztály	Pozitív	Valós pozitív	Hamis negatív
	Negatív	Hamis pozitív	Valós negatív

III. Leíró statisztikák

Általános mutatók

A kutatás során használt korpusz a szűrési lépések elvégzése után 12 800 kutatási absztraktot tartalmaz. Ezek a prediktált változó alapján egyenlő arányban oszlanak meg, így végül 6400 ember által írt, valamint 6400 mesterséges intelligencia által generált szöveg került felhasználásra. Az előfeldolgozási lépések elvégzése után az ember osztályába tartozó szövegek átlagosan 116 token hosszúságúak, míg az AI absztraktok ettől valamelyest elmaradnak. A tokenek átlagos karakterszáma a mesterséges intelligencia által generált absztraktok esetében magasabb, míg a főnevek és az igék aránya megközelítőleg hasonló a két osztályban.

	Ember	AI
Tokenek száma	116	105
Tokenek hossza karakterben	6.91	7.36
Főnevek aránya	0.56	0.58
Igék aránya	0.10	0.11

1. Táblázat – Általános mutatók

Olvashatósági és szókincsgazdagsági mutatók

Az alábbi táblázatban a két osztály átlagos olvashatóságértékei láthatóak. Welch-féle T próbát alkalmazva elmondható, hogy minden mutató mentén statisztikailag szignifikáns különbségek figyelhetők meg. A Flesch Reading Ease értékek alapján könnyebben értelmezhetőek az ember által írt absztraktok. A Flesch-Kincaid Grade Level és a Smog Index mutatók alapján az AI generálta szövegek megértéséhez több elvégzett iskolai év szükséges. Ezekkel némileg ellent mondanak a Dale-Chall Readability dimenzión felvett értékek, melyek szerint az ember írta szövegek több nehezen értelmezhető szót tartalmaznak.

	Ember	AI	p-érték (Welch-féle T)
Flesch Reading Ease	27.98	19.51	< 0.01
Flesch-Kincaid Grade Level	15.10	16	< 0.01
Smog Index	15.80	16.96	< 0.01
Dale-Chall Readability Score	10.76	10.59	< 0.01

2. Táblázat - Olvashatósági mutatók

A szókincsgazdagsági mutatók mentén szintén statisztikailag szignifikáns különbségek mutatkoznak az ember által írt, és a mesterséges intelligencia által generált szövegek között. Az RTTR (Root Type-Token Ratio) és a Herdan féle C érték alapján elmondható, hogy az ember írta absztraktok valamelyest magasabb egyedi szó aránnyal rendelkeznek. Ezzel egy irányba mutatnak a Yule féle K és Simpson féle D mérőszámokon felvett értékek is, melyek szerint az AI generálta szövegekre minimálisan magasabb repetitivitás jellemző.

	Ember	AI	p-érték (Welch-féle T)
Root Type-Token Ratio	7.36	7.03	< 0.01
Yule’s K	98.84	109.38	< 0.01
Simpson’s D	0.010	0.011	< 0.01
Herdan’s C	0.930	0.928	< 0.01

3. Táblázat - Szókincsgazdagsági mutatók

Leggyakoribb n-gram-ok

Az alábbi táblázatokban a teljes korpusz, valamint a két vizsgált osztály leggyakoribb uni- és bigramjai, és azok abszolút gyakoriságértékei szerepelnek. Megfigyelhető, hogy az AI osztály esetében magasabbak ezek a gyakoriságok, amely arra utalhat, hogy a generálás során

használt nagy nyelvi modell hajlamos a jól bevált, ismétlődő szerkezetek használatára. Mindkét osztály esetében igaz, hogy a leggyakoribb unigramok között többnyire általános, tudományos kontextushoz kapcsolódó szavak jelennek meg gyakran (például „study”, „model”, „result”). A bigramokat illetően megfigyelhetők bizonyos különbségek a két osztály között. Míg az AI generálta szövegek esetében elsősorban sablonszerű, általánosabb szókapcsolatok szerepelnek a leggyakoribb 10 között, addig az ember által írt absztraktoknál inkább a tudományos szakkifejezések kerülnek előtérbe. Az AI osztály esetében magasan kiemelkedik a „paper present” bigram, mely a mesterséges intelligencia szerzőségének erős ismertetőjegye lehet. Érdekes továbbá megfigyelni, hogy a szövegközi hivatkozásoknál használt „et al” kifejezés az ember által írt absztraktok esetén a legtöbbször előforduló szókapcsolatok egyikeként jelenik meg.

	Teljes korpusz		Ember		AI	
1.	use	11327	model	5664	study	6976
2.	model	10465	use	5442	result	6168
3.	result	9760	result	3592	use	5885
4.	study	9586	data	3113	provide	5386
5.	provide	7065	star	2769	model	4801
6.	method	7001	method	2764	paper	4563
7.	data	6561	study	2610	method	4237
8.	galaxy	6297	galaxy	2430	approach	4141
9.	paper	6292	present	2272	galaxy	3867
10.	star	6260	field	2256	property	3805

4. Táblázat - Leggyakoribb unigramok

	Teljes korpusz		Ember		AI	
1.	paper present	1500	star formation	482	paper present	1326
2.	star formation	1342	black hole	448	star formation	860
3.	black hole	1212	magnetic field	432	black hole	764
4.	magnetic field	1075	et al	369	propose method	703
5.	propose method	862	neural network	366	paper investigate	688
6.	paper propose	831	paper propose	197	magnetic field	643
7.	neural network	769	stellar mass	197	paper propose	634
8.	paper investigate	734	dark matter	195	provide insight	632
9.	result demonstrate	684	light curve	194	formation evolution	619
10.	provide insight	665	paper present	174	result demonstrate	610

5. Táblázat - Leggyakoribb bigramok

IV. Eredmények

IV.1 N-gram alapú klasszifikációs modellek

Kutatási kérdésem megválaszolása érdekében n-gram vektorizációt használva tanítottam klasszifikációs modelleket. Összesen három algoritmust, logisztikus regressziót, Naive Bayes-t, és XGBoost-ot használtam, melyeket két különböző vektorizációs módszerrel (CountVectorizer, TfidfVectorizer) kombináltam. Az egyes vektorizációkat többféle n-gramot alapul véve is elvégeztem, külön vizsgáltam unigramokat és bigramokat, majd pedig ezeket vegyesen. A vektorok dimenzióinak számát több módon is szabályoztam. A vektorizációs függvények min_df paraméterét 20-ra állítottam, azaz azokat az n-gramokat, amelyek nem szerepeltek legalább 20 dokumentumban, eltávolítottam. Emellett a max_df paraméter 0.8-as értékre történő beállításával azokat az n-gramokat is kiszűrtem, amelyek a szövegek több, mint 80%-ban megjelentek. Annak érdekében, hogy az egyes algoritmusok esetén megtaláljam a legideálisabb vektorizációs eljárást, minden kombinációval 5-fold keresztvalidációt végeztem a tanulóhalmazon belül, melynek eredményei a 3. számú táblázatban láthatóak. Ezt megelőzően az adatot 80/20 arányban osztottam tanuló- és teszthalmazra a scikit-learn modul train_test_split [33] függvényének segítségével.

Minden klasszifikációs algoritmus és vektorizációs eljárás kombinációjára igaz, hogy kifejezetten magas pontossággal prediktáltak a modellek, az „accuracy" értékek szórásai pedig kivétel nélkül 1% alattiak, mely arra utal, hogy nem áll fent a túlillesztés veszélye, hiszen nem volt jelentősége annak, hogy melyik szövegek kerültek az egyes fold-okba. Összességében elmondható, hogy a leghatékonyabb vektorizációs módszernek az tűnik, amely az uni- és bigramok súlyozatlan gyakoriságait kombinálja, ugyanis minden klasszifikációs algoritmus ezekkel a változókkal érte el a legmagasabb pontosság értéket. A Naive Bayes modellek esetén figyelhetők meg a másik két algoritmushoz képest relatíve „alacsony", 90% alatti „accuracy" értékek.

Bár alapértelmezett paraméterekkel futtattam a modelleket, a magas pontosságértékek miatt nem éreztem szükségesnek azok további finomhangolását.

[33] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

		Logisztikus regresszió		Naive Bayes		XGBoost	
		átlagos pontosság	szórás	átlagos pontosság	szórás	átlagos pontosság	szórás
Count	unigram	0.942	0.003	0.919	0.006	0.951	0.004
	bigram	0.934	0.006	0.889	0.008	0.917	0.007
	uni és bigram	**0.954**	0.003	**0.946**	0.004	**0.952**	0.003
TF-IDF	unigram	0.946	0.004	0.901	0.009	0.948	0.946
	bigram	0.929	0.006	0.887	0.006	0.913	0.004
	uni és bigram	0.953	0.004	0.939	0.005	0.949	0.003

6. Táblázat - Tanulóhalmazon mért 5-fold keresztvalidációs átlagos pontosság értékek és szórások

A validációs fázis után a mindhárom klasszifikációs algoritmus esetén kiválasztottam a legmagasabb predikciós pontosságot elérő vektorizációt, majd az így kapott modelleket a tanulóhalmaz teljes egészén újratanítottam annak érdekében, hogy az addig nem használt különálló teszthalmazon is megmérjem teljesítményüket. A keresztvalidációs eredményekhez hasonlóan elmondható, hogy a modellek a teszthalmazon is kiemelkedő pontosság értékeket értek el, azaz az esetek túlnyomó többségében képesek voltak helyes döntést hozni a kutatási absztraktok szerzőit illetően. A legjobb teljesítményre a logisztikus regressziós modell volt képes, 95.8%-os predikciós pontossággal.

	Pontosság	F1	Precízió	Visszahívás
Logisztikus regresszió	0.958	0.958	0.964	0.952
Naive Bayes	0.941	0.942	0.937	0.947
XGboost	0.956	0.955	0.962	0.948

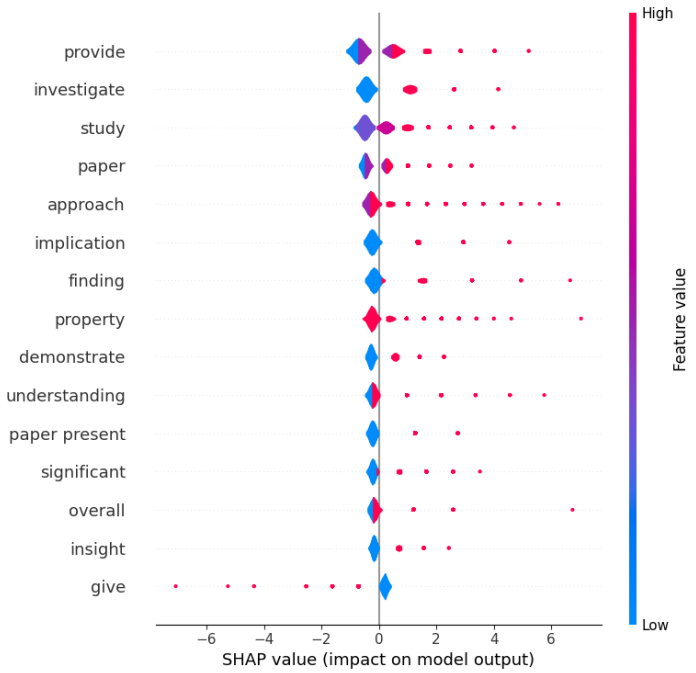
7. Táblázat – Klasszifikációs modellek teszthalmazon mért eredményei

A teszthalmazba tartozó 2560 szövegből 2322 (91%) esetén hozott helyes döntést mind a három modell, 42 (2%) absztraktról pedig éppen az ellenkezője mondható el. 196 (7%) teszt szöveg esetében nem volt egyetértés a modellek között (lásd 3. és 4. Melléklet).

IV.2 Változók fontossága

Uni- és bigramok

Elsőként a legmagasabb pontosságot elérő modell esetében vizsgáltam meg (logisztikus regresszió uni-és bigramokkal), hogy mely változók hatnak leginkább a predikciós döntésre, azaz mely n-gramok gyakoriságértékei játsszák a legnagyobb szerepet abban, hogy a modell emberi szerzőnek vagy a ChatGPT-nek tulajdonítja az adott absztraktot. A változók fontosságainak vizsgálata céjából SHAP (SHapley Additive exPlanations) értékeket [34] számítottam. Ezek az értékek a magyarázó változók predikcióra való hatását számszerűsítik. A SHAP módszer előnye, hogy bármilyen gépi tanulási modell esetében alkalmazható, ezáltal egységesen vizsgálhatók a változók fontosságai olyan esetekben is, amikor több modellt alkalmazunk (Shah és munkatársai, 2023). A 3. ábrán fentről lefelé sorrendben a 15 legfontosabb szerepet játszó n-gram látható, SHAP értékeik alapján ezeknek az uni- és bigramoknak volt a legnagyobb ráhatása a predikciós döntésre.



3. Ábra – Logisztikus regressziós modell 15 legfontosabb uni- és bigramja

A két legnagyobb jelentőséggel bíró szónak a „provide" és az „investigate" mutatkozik, ezek jelenléte erősen az AI irányába mozdítja a predikciót, mely abból látható, hogy a piros pontok az x-tengelyen a 0 értéktől jobbra helyezkednek el. Ellenkező esetekben, amikor ezek a szavak

[34] https://shap.readthedocs.io/en/latest/

nincsenek jelen a szövegben (kék foltok), valamelyest az ember irányába tolódik a prediktált osztály. Ebből arra lehet következtetni, hogy az AI generálta absztraktok gyakorta használnak olyan sablonszerű szókapcsolatokat melyek tartalmazzák ezeket a szavakat (például „this study investigates").

Hasonló szereppel bírnak a „study", „paper", „implication" és „finding" szavak is, melyek jelenléte pozitívan hat az AI osztályba tartozás valószínűségére. A konzisztensen jobb oldalon elhelyezkedő piros pontok azt szemléltetik, hogy a modell ezeket a szavakat a mesterséges intelligenciával asszociálja. Ezen unigramok jelenléte a korpuszban nem meglepő, hiszen szorosan kapcsolódnak az akadémiai szóhasználathoz, azonban gyakori, repetitív használatuk erős AI-ra utaló prediktív erővel bírhat.

Az „approach" és „property" szavak bár szintén jelentősen befolyásolják a predikciót SHAP értékeik alapján, de erősen kontextusfüggő, hogy az AI vagy pedig az ember irányába mozdítják-e el a döntést. Ez onnan látható, hogy az x-tengely 0 értéktől jobbra, és a 0 érték közelében egyaránt piros színű pontok helyezkednek el.

Egyedüli bigramként a „paper present" is a legjelentősebb változók egyikeként jelenik meg SHAP értéke alapján. Jelenléte az AI szerzőségére utal, ha pedig nem tartalmazza ezt a szókapcsolatot az adott szöveg, az bár gyengén, de az ember irányába mozdítja a predikciót. A „paper present" szókapcsolat magas jelentőségértéke is arra utal, hogy jellemző a ChatGPT-re bizonyos sablonszerű kifejezések használata.

A Naive Bayes modell esetében elmondható, hogy a legfontosabb n-gramok (lásd 1. Melléklet) nagy átfedésben vannak az imént bemutatott logisztikus regressziós modell legjelentősebb változóival. Hasonlóan nagy szerep jut például a „provide", „study", „paper" vagy „implication" szavaknak mindkét algoritmus predikciói során, valamint abban is hasonlítanak, hogy a legfontosabb n-gramok jelenléte többnyire az AI irányába mozdítja a döntést. Az XGBoost modell esetében azonban jóval alacsonyabb az átfedés, más szavak játszanak fontos szerepet az ember és AI közötti osztályozás során (lásd 2. Melléklet). Érdekes megfigyelni, hogy számos, tudományos szövegekben használatos, latin eredetű rövidítés is nagy hatású változóként

szerepel az XGBoost modellben - ilyen például az „eg” (exempli gratia), „ie”[35] (id est), vagy az „et”, mely az „et al” (et alii) szókapcsolatból ered.

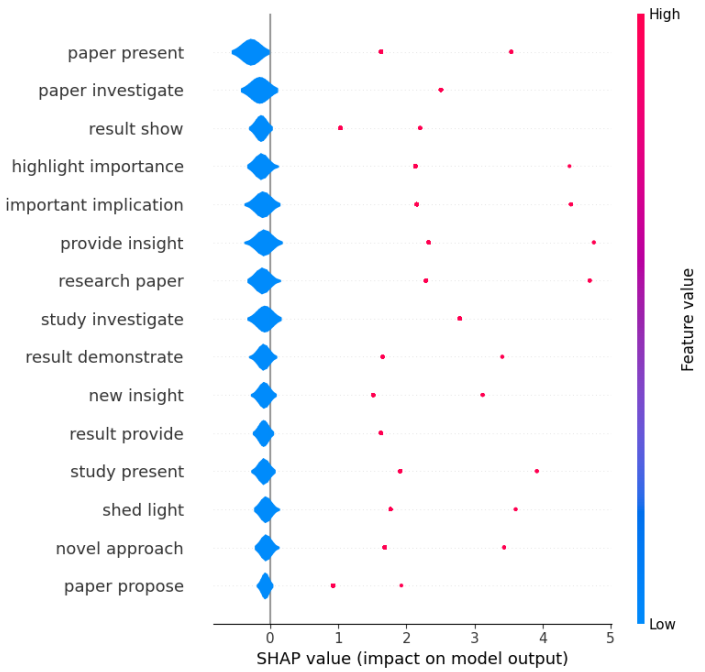
Bigramok

Annak érdekében, hogy kifejezetten a szókapcsolatok fontosságát tudjam vizsgálni, az egyik bigram gyakoriság értékeken tanított modellem (logisztikus regresszió) esetében is elvégeztem az imént bemutatott SHAP elemzést.

A 15 legnagyobb jelentőséggel bíró bigram mindegyikéről elmondható, hogy jelenlétük a szövegben pozitív irányba mozdítja a predikciót, azaz ezen szókapcsolatok előfordulása az AI szerzőségére utal. Ez onnan látható, hogy minden piros pont az x-tengely 0 értékétől jobbra helyezkedik el. Ellentétes esetben, tehát amikor ezek a bigramok nem fordulnak elő a szövegben, az gyenge negatív vagy semleges hatással van a predikcióra.

Az ábrán jól látható, hogy bizonyos tokenpárok adott esetben kifejezetten magas SHAP értékeket is elérnek (4-5), ilyenek például a „highlight importance”, „important implication”, „provide insight” vagy akár a „research paper”. Ez arra utal, hogy a vizsgált nagy nyelvi modell gyakran használ ilyen sablonszerű formális, akadémia nyelvezetbe illeszkedő szópárokat, melyek alapján az általam tanított osztályozó egyszerűen felismeri, hogy az adott absztrakt a mesterséges intelligenciától származik. Továbbá azt is egyértelműen mutatják az alábbi SHAP értékek, hogy alapvetően az AI osztály valószínűségére pozitív hatással lévő változók a legdominánsabbak a predikciós döntés meghozatalakor, az embertől származó absztraktok esetén nem ismerhetők fel ennyire erős karakterisztikák.

[35] tisztítás előtt e.g. és i.e.



4. Ábra – Logisztikus regressziós modell 15 legfontosabb bigramja

IV.3 Predikció a transzformer alapú detektorral

Szakdolgozatomban a klasszikus gépi tanulási klasszifikációs algoritmusok mellett viszonyítási alapként egy már betanítva elérhető nyílt forráskódú transzformer architektúrájú detektort is alkalmaztam. A modell minden bemenetre egy 0 és 1 közötti értékkel válaszol, amely annak valószínűségét mutatja meg, hogy az adott szöveget valamelyik nagy nyelvi modell generálta. A modell egyébiránt lokálisan futtatható korlátozás nélkül, amely nagyban megkönnyítette az automatizált detekció folyamatát. Először a standard 0.5-ös döntési küszöb mellett értékeltem ki a predikciókat ugyanazon a teszthalmazon, amelyet a klasszikus algoritmusoknál is használtam. Az eredmények a 8. táblázatban láthatóak.

Pontosság	Precízió	Visszahívás	F1
0.83	0.75	1	0.85

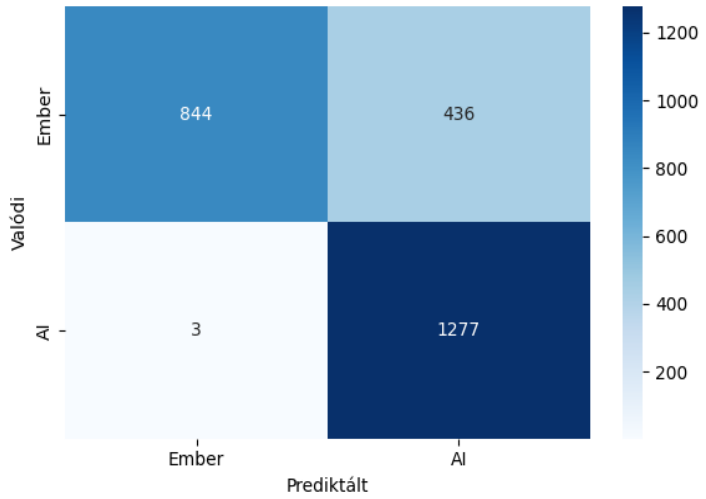
8. Táblázat – Transzformer alapú detektor teszthalmazon mért eredményei

A modell összességében 83%-os pontossággal prediktált, amely első olvasatra relatíve jó teljesítménynek mondható. Ha azonban részletesebben megvizsgáljuk a kapott eredményeket, a két osztály (ember és AI) detektálási hatékonyságát illetően jelentős különbségek figyelhetők

meg. A detektor a mesterséges intelligencia által generált szövegek felismerésében kiemelkedően teljesített, az 1-es visszahívás érték arra utal, hogy az AI által készített absztraktok közel 100%-át [36] sorolta a megfelelő osztályba. A precízió érték alapján az AI címkével ellátott szövegek háromnegyede ténylegesen mesterséges intelligencia eredetű volt.

Az ember által írt absztraktok esetén ennél gyengébb teljesítményt mutatott a használt detektor. Bár majdnem minden emberiként címkézett absztrakt ténylegesen valamilyen emberi szerzőtől származott, a modell az ember által írt szövegek egyharmadát tévesen AI eredetűnek minősítette (lásd 5.ábra).

Az alábbi ábrán látható konfúziós mátrix még jobban szemlélteti ezt az eltolódást: 844 ember által írt absztraktot sikerült helyesen besorolni, ezzel szemben 436 esetben volt téves a predikció. A másik osztály eredményeit tekintve jól látható, hogy rendkívül ritka az, amikor a detektor AI generált tartalmat tévesen emberi szerzőnek tulajdonít.



5. Ábra – Transzformer alapú detektor konfúziós mátrixa 0.5-ös küszöbértékkel

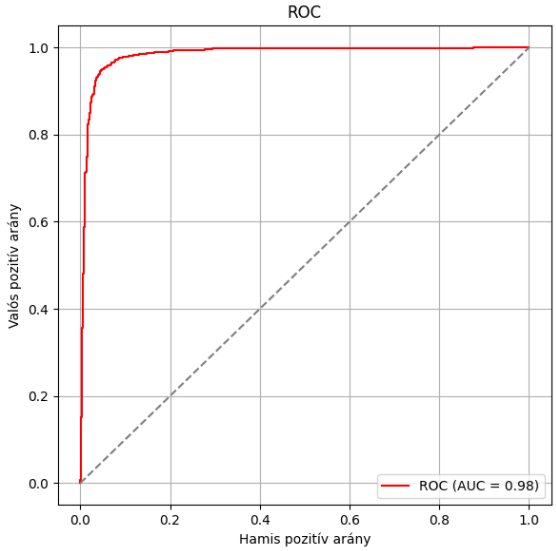
A 0.5-ös küszöbérték mellett megfigyelt eredmények arra utalnak, hogy a modell elsődlegesen a mesterséges intelligencia által generált szövegek kiszűrésére van optimalizálva, más szóval elsődleges fókusz a hamis negatív eredmények minimalizálása. Ez a tulajdonság kifejezetten hasznos lehet olyan szituációkban, amikor egy szöveg hitelességének megítélése szempontjából kiemelten fontos, hogy az esetleges AI generált tartalmat azonosítani tudjuk. Másfelől viszont látni kell, hogy a kimagaslóan pontos detekciónak ára van, ugyanis így a hamis

[36] a klasszifikációs eredmények 2 tizedesjegyre vannak kerekítve

pozitív esetek is nagyobb arányban lesznek jelen, mely komoly kérdéseket vet fel az ilyen detektorok akadémiai közegben való alkalmazhatóságát illetően.

A detektor teljesítményének átfogó értékeléséhez ROC (Receiver-operating characteristic curve) görbét alkalmaztam, amely azt mutatja meg, hogy különböző küszöbértékek mellett hogyan alakul a hamis pozítv, és valós pozitív predikciók aránya. A görbe bal felső sarokhoz közeli csúcsosodása azt jelzi, hogy létezik olyan ideális küszöbérték tartomány, ahol a két vizsgált arányszám egyensúlyba hozható, azaz alacsony hamis pozitív arány mellett magas valódi pozitív arány érhető el.

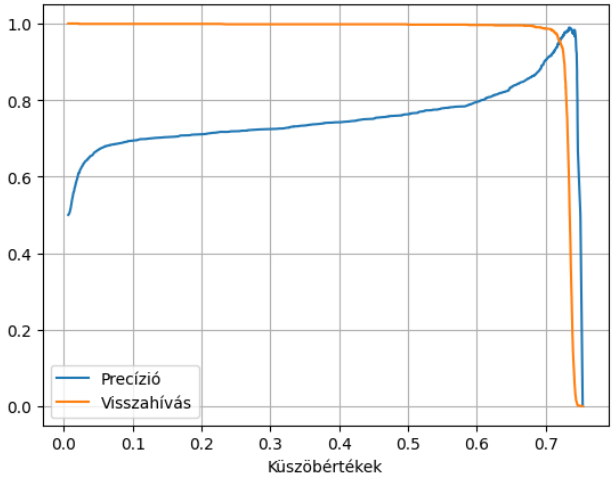
Az AUC (area under curve) érték egy olyan mérőszám, mely az imént említett csúcsosodás mértékét számszerűsíti, azaz a ROC görbe alatti terület nagyságát mutatja meg. Az érték a transzformer alapú detektor esetén közel maximális (0.98), azaz a küszöbértéktől független osztályozási teljesítménye kiemelkedően jónak mondható.



6. Ábra – Transzformer alapú detektor ROC görbe 0.5-ös küszöbértékkel

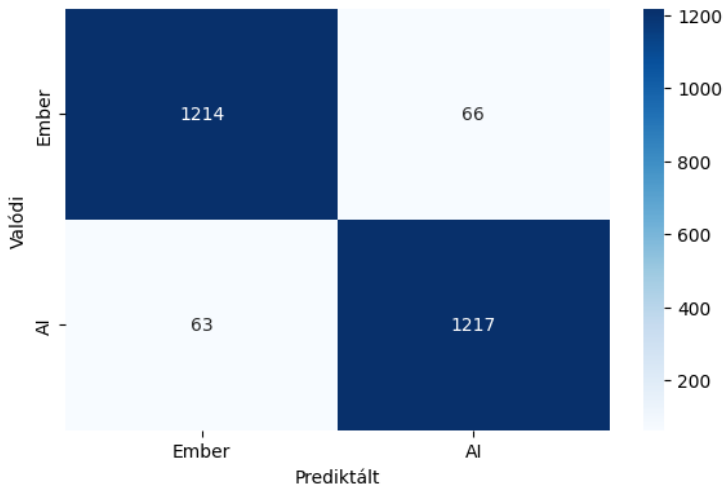
Jelen esetben releváns cél lehet megtalálni azt a küszöbértéket, mely mellett a modell úgy képes relatíve pontosan osztályozni, hogy a hamis pozitív predikciók számát minimalizálja. Ennek oka, hogy kutatásom kontextusában ezek a hamis pozitív predikciók tulajdonképpen azt jelentik, hogy egy ember munkájáról tévesen azt állítjuk, hogy az valójában a mesterséges intelligencia műve. Ezen küszöbérték megtalálásában segítséget nyújthat a lentebb látható precizitás-visszahívás görbe, melyet egy teszthalmazon kívüli validációs halmaz szövegeinek

prediktált valószínűségei alapján készítettem el. Mivel cél most a precizitás értékének maximalizálása, azaz annak elérése, hogy az AI osztályba sorolt absztraktok magas százaléka ténylegesen AI generált legyen, azt a pontot kell megtalálni, ahol ez a mérőszám kellően magas, de a visszahívás értéke még nem csökken túlzottan alacsony szint alá. Az ábra alapján a keresett küszöbérték 0.72 környékén található meg.



7. Ábra – Transzformer alapú detektor precizitás-visszahívás görbéje validációs halmazon

Az alábbi ábrán az új, 0.72-es küszöbértékkel kalkulált konfúziós mátrix látható. Bár valamelyest növekedett a hamis negatív predikciók száma, azaz a szó szoros értelmében vett detekciós képesség romlott, a hamis pozitív esetek gyakorisága olyan mértékben esett vissza, hogy a modell teljesítménye összességében nagy mértékben javult. A predikciós pontosság 0.83-ról 0.95-re emelkedett.



8. Ábra – Transzformer alapú detektor konfúziós mátrixa optimalizált küszöbértékkel

IV.4 Kitekintés – Klasszifikációs modellek tesztelése más nagy nyelvi modellek szövegeivel

Mivel a kutatásom során használt absztraktok AI verziónak 100%-át egy bizonyos LLM készítette (ChatGPT 3.5 turbo), felmerült bennem az a kérdés, hogy a kifejezetten ChatGPT generálta szövegeken tanított klasszifikációs modellek vajon más nagy nyelvi modellek output-jait is képesek lehetnek-e detektálni. Más szóval létezhetnek olyan mesterséges intelligenciára jellemző általános szövegalkotási tulajdonságok, melyek alapján modelltől függetlenül beazonosítható az AI szerzősége? Lehetséges-e, hogy például a ChatGPT és a Claude olyan hasonló módon generál szöveget, hogy a kizárólag az egyiket „ismerő" detektorok a másikat is felismerik? Kutatásom végső fázisában azzal foglalkoztam, hogy ezekre a kérdésekre megpróbáljak választ adni.

Mindezek érdekében az eredetileg használt teszthalmazban található ember által írt eredeti kutatási absztraktok egy véletlenszerűen megválasztott részhalmazának szövegeihez (szám szerint 30 darabhoz) több különböző nagy nyelvi modellel generáltattam AI verziót. A generálási folyamatot limitálta, hogy a legnépszerűbb modellek esetében általában elmondható, hogy csak a weboldalaik felhasználói felületén keresztül elérhető verziók férhetők hozzá szabadon, az API verziók használata legtöbbször korlátozott, emiatt a generálás nagyrészt manuálisan történt. Fontos azonban hozzátenni, hogy bizonyos nyílt forráskódú modellek lokálisan, úgynevezett „on-device" módon is futtathatók, melyet például a Meta által fejlesztett Llama modell esetében ki tudtam használni. Ennek megvalósításához az Ollama[37] nevű nyílt forráskódú programot használtam.

Az absztraktok generálásához használt prompt megegyezett az eredeti adatbázis létrehozásánál használt prompt-tal, a választott modellek pontos verziói pedig a következők voltak: Claude 3.5 Haiku, DeepSeek V3, Gemini 2.0 Flash, Llama 3.2. Mivel ebben az esetben most minden teszt szöveg ugyanabba az osztályba tartozott, a kiértékelést a visszahívás érték mentén végeztem, azaz minden LLM esetén azt vizsgáltam, hogy a 30 mesterséges intelligencia által generált absztrakt mekkora részét sorolják ténylegesen az AI osztályba modelljeim.

	Logisztikus Regresszió	Naive Bayes	XGBoost
Claude 3.5 Haiku	1	1	1

[37] https://github.com/ollama/ollama

DeepSeek V3	0.63	0.37	0.70
Gemini 2.0 Flash	0.97	0.97	0.83
Llama 3.2.	0.93	0.93	0.90

9. Táblázat – Más nagy nyelvi modellek visszahívás értékei

A Claude 3.5 Haiku modellverzió esetében mindhárom algoritmus 100%-os visszahívást ért el, ami arra utalhat, hogy a Claude szöveggenerálási stílusa hasonlíthat a ChatGPT-éhez, legalábbis tudományos, formális szövegek esetében.

Ezzel szemben a DeepSeek által generált absztraktokat már jóval nehezebben ismerték fel az osztályozók, különösen igaz ez a Naive Bayes-re, amely mindössze 0.37-es visszahívást értéket ért el. Ez arra utalhat, hogy a DeepSeek szövegalkotási specifikumai eltérnek a ChatGPT esetében felismerhető AI mintázatoktól, vagy olyan módon struktúrálják a szövegeket, amely kevésbé illeszkedik a klasszifikáló modellek által felismert jellegzetességekhez.

A Gemini 2.0 Flash és a Llama 3.2 esetében minden osztályozó legalább 0.9 körüli, de inkább feletti visszahívás értéket ért el, amiből szintén arra lehet következtetni hogy ebben a specifikus kontextusban hasonló nyelvi mintázatok alapján alkothatnak szöveget az imént említett modellek és a tanítás során használt ChatGPT verzió.

Ezek az eredmények azt sugallhatják, hogy létezhetnek általános nem modell-specifikus mesterséges intelligenciára jellemző szövegalkotási mintázatok. Fontos azonban újra megemlíteni, hogy az imént bemutatott eredmények kifejezetten alacsony mintaelemszámok mellett születtek meg, ennek okán elemzésem ezen része elsősorban kitekintésként, egyfajta potenciális jövőbeli kutatási irány felvázolásaként értelmezendő.

V. Összegzés

V.1 Diszkusszió

Kutatásom során azt vizsgáltam, hogy egyes hagyományos gépi tanulási algoritmusok milyen hatékonysággal képesek mesterséges intelligencia által generált szöveget detektálni egy, modern transzformer alapú detektorhoz képest.

Elsőként, az egyes szókincsre és olvashatóságra vonatkozó mérőszámok elemzése azt mutatta, hogy statisztikailag szignifikáns különbségek figyelhetők meg a mesterséges intelligencia által

generált, és az emberi szerzők írta szövegek között. Összességében elmondható, hogy az ember által készített absztraktok valamelyest könnyebben olvashatóak és értelmezhetőek, míg az AI-ra jellemzőbb a kevésbé változatos, repetitív szóhasználat, utóbbit a leggyakoribb n-gramok vizsgálata is megerősítette. Ezek alapján az a következtetés vonható le, hogy bár a kutatás során alkalmazott nagy nyelvi modell hatékonyan imitálja a tudományos nyelvezetet, gyakran használ olyan sablonos szófordulatokat, melyek megkönnyíthetik szerzőségének beazonosítását.

A klasszikus gépi tanulási modellekről elmondható, hogy magas pontossággal képesek az emberi és gépi szerzőtől származó absztraktokat klasszifikálni. A legjobb vektorizációs technikának az uni- és bigramok súlyozatlan gyakoriságértékeit együttesen felhasználó vektorizáció bizonyoult, az algoritmusokat illetően pedig a logisztikus regresszió érte el a legjobb teljesítményt (95,8%-os pontosság). A változók fontosságának vizsgálatára irányuló SHAP elemzés kimutatta, hogy bizonyos n-gramok jelenléte számottevő hatással van a predikciós döntésre, elsősorban az AI osztály irányába mozdítva a döntést, ilyen kifejezések például a „provide” az „investigate” a „study” vagy a „paper present”. Ezek az eredmények is azt a feltételezést erősíthetik, hogy a nagy nyelvi modellek által generált szövegek jól azonosítható, visszatérő mintázatokat tartalmaznak, melyekre a klasszikus gépi tanulási algoritmusok is képesek „rátanulni”.

A transzformer alapú detektor esetén az volt megfigyelhető, hogy az általam alapértelmezettként használt 0.5-ös küszöbérték mellett rendkívül magas pontossággal azonosította a mesterséges intelligencia által generált tartalmat, ez viszont magas hamis pozitív aránnyal járt együtt. Egy új, optimalizált küszöbérték mellett a transzformer alapú detektor is 95%-os pontossággal címkézte a teszthalmaz szövegeit, továbbá a hamis pozitív predikciók aránya is jelentősen csökkent.

Kutatásom végső szakaszában azt vizsgáltam, hogy a kizárólag ChatGPT által generált szövegeken tanított modelljeim képesek-e más nagy nyelvi modellek tartalmait detektálni. Eredményeim azt mutatják, hogy bizonyos modellek, mint például a Claude 3.5 Haiku szerzőségei szinte tökéletes pontossággal azonosíthatóak voltak, míg más LLM-ek, mint például a DeepSeek V3, esetén kevésbé volt hatékony a detekció. Ebből az a következtetés vonható le, hogy létezhetnek általános, nem modellspecifikus szövegalkotási jellemzők, melyek alapján

modelltől függetlenül detektálható az AI generálta tartalom, ez azonban további, nagyobb mintaelemszámon végzett vizsgálatot igényel.

Összességében elmondható, hogy a klasszikus gépi tanulási módszerek hasonlóan magas predikciós pontosságot képesek elérni, mint a vizsgált transzformer alapú detektor.

V.II. Limitációk, jövőbeli irányok

Bár az imént bemutatott eredmények azt sugallják, hogy a klasszikus gépi tanulási klasszifikációs algoritmusok az újabban sokkal inkább elterjedt transzformer architektúrájú detektorok alternatívái lehetnek, bizonyos tények árnyalhatják ezt a kutatási kérdésem szempontjából pozitív képet.

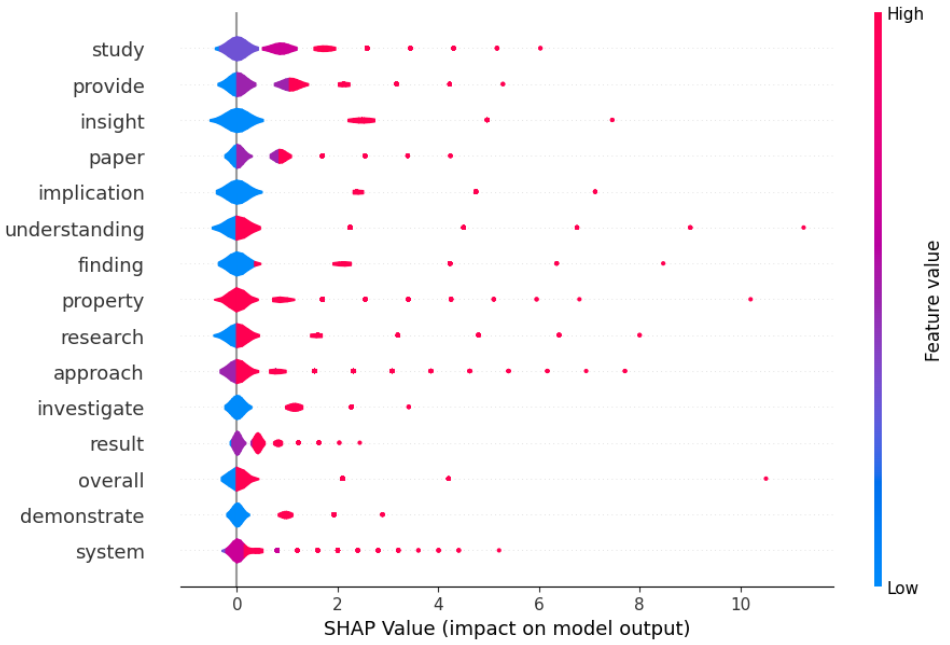
A klasszifikációs modellek tanítására használt absztraktok AI verzióinak 100 százalékát egy bizonyos nagy nyelvi modell generálta (ChatGPT 3.5 Turbo), ami azt jelenti, hogy az algoritmusok nem feltétlenül általános AI-ra jellemző szövegalkotási specifikumokra tanultak rá, hanem elsősorban az említett modell mintázataira. Bár kutatásom végső fázisában foglalkoztam ezzel a problémával, az ott kapott eredményeim az alacsony elemszámok miatt elsősorban egy lehetséges jövőbeli kutatási irány kijelöléseként értelmezendők.

Fontos továbbá leszögezni, hogy az a tény, hogy az általam használt klasszifikációs algoritmusok magas pontossággal prediktáltak, nem azt jelenti, hogy általánosságban bármilyen típusú szöveg esetén hatékony detektorként működnének. Modelljeim specifikusan kutatási absztraktok szövegeivel lettek tanítva, amiből az következik, hogy elsősorban ugyanilyen típusú szövegek esetén alkalmasak az AI szerzőségének azonosítására. A viszonyítási alapként kezelt transzformer alapú detektorról ez azonban nem mondható el, mivel annak tanítása során változatos forrásokból származó szövegeket vettek alapul a készítők. Jövőbeli kutatási irányként merül fel ennek okán annak vizsgálata is, hogy hagyományos gépi tanulási algoritmusok képesek lehetnek-e szövegtípustól függetlenül hatékonyan működő AI detektorként funkcionálni.

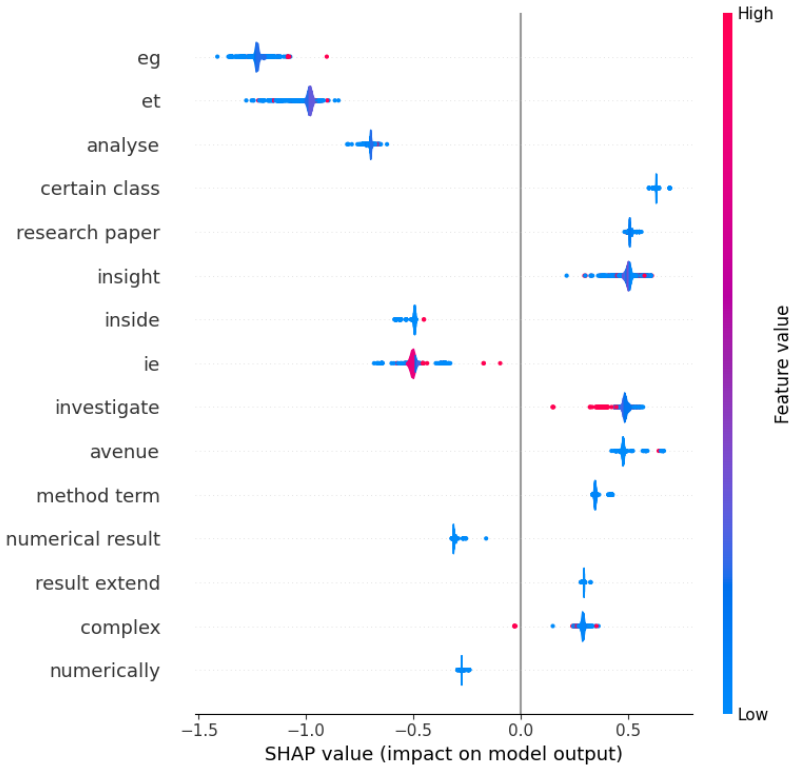
Kutatási eredményeim továbbá annak fényében értelmezendők, hogy az absztraktok AI verzióinak generálásánál használt prompt hatása nem lett vizsgálva. Elképzelhető, hogy más prompt alapján generált szövegek esetén kevésbé pontos detekcióra lennének képesek az alkalmazott algoritmusok. A kutatás érdekes jövőbeli iránya lehet például annak vizsgálata is,

hogy specifikus prompt-ok használatával létrehozhatóak-e olyan szövegek, melyek esetén az AI szerzősége kifejezetten nehezen azonosítható be.

Mellékletek:



1. Melléklet – Naive Bayes modell 15 legfontosabb uni- és bigramja



2. Melléklet – XGBoost modell 15 legfontosabb uni- és bigramja

3. Melléklet – Példák mindhárom modell által helytelenül címkézett szövegekre

1. Valódi osztály: AI

This paper reports on the observation of two distinct phenomena in proton-proton scattering at high energies, namely the "break" and the "dip." These structures manifest themselves as anomalies in the diffraction cone, which is the angular distribution of scattered particles that emerge close to the incident beam. We present a theoretical model that explains these features as arising from the quantum mechanics of the interaction between the colliding protons. The model makes specific predictions for the energy and angular dependence of the break and dip, which we test against new experimental data from the Large Hadron Collider

2. Valódi osztály: Ember

We present a theoretical study on the dynamical properties of three-dimensional arrays of Josephson junctions. Our results indicate that such superconducting networks represent highly sensitive 3D-SQUIDs having some major advantages in comparison with conventional planar SQUIDs. The voltage response function of 3D-SQUIDs is directly related to the vector-character of external electromagnetic fields. The theory developed here allows the three-dimensional reconstruction of a detected external field including phase information about the field variables. Applications include the design of novel magnetometers, gradiometers and particle detectors.

4. Melléklet – Példák olyan szövegekre, melyeket eltérően címkéztek a modellek

1. Valódi osztály: AI

This paper investigates the long-time asymptotic behavior of solutions to the fifth-order modified Korteweg-de Vries (mKdV) equation. Using the linearized stability theory and a modified energy method, we prove that all solutions to the mKdV equation approach a finite number of solitary waves as t goes to infinity. We also provide a description of the convergence rate, showing that it is at least exponential when the solitary waves are well separated. The results obtained in this paper generalize the existing theory on the asymptotic behavior of solutions to the KdV equation.

2. Valódi osztály: Ember

Deep neural networks can be powerful tools, but require careful application-specific design to ensure that the most informative relationships in the data are learnable. In this paper, we apply deep neural networks to the nonlinear spatiotemporal physics problem of vehicle traffic dynamics. We consider problems of estimating macroscopic quantities (e.g., the queue at an intersection) at a lane level. First-principles modeling at the lane scale has been a challenge due to complexities in modeling social behaviors like lane changes, and those behaviors' resultant macro-scale effects. Following domain knowledge that upstream/downstream lanes and neighboring lanes affect each others' traffic flows in distinct ways, we apply a form of neural attention that allows the neural network layers to aggregate information from different lanes in different manners. Using a microscopic traffic simulator as a testbed, we obtain results showing that an attentional neural network model can use information from nearby lanes to improve predictions, and, that explicitly encoding the lane-to-lane relationship types significantly improves performance. We also demonstrate the transfer of our learned neural network to a more complex road network, discuss how its performance degradation may be attributable to new traffic behaviors induced by increased topological complexity, and motivate learning dynamics models from many road network topologies..

Felhasznált irodalom:

Abubakar, H. D., Huspi, S. H., & Umar, M. (2022). A Scheme of Pairwise Feature Combinations to Improve Sentiment Classification Using Book Review Dataset. *International Journal of Innovative Computing*, *12*(1), 25-33.

André, C. M., Eriksen, H. F., Jakobsen, E. J., Mingolla, L. C., & Thomsen, N. B. (2023). Detecting AI authorship: Analyzing descriptive features for AI detection. In *7th Workshop on Natural Language for Artificial Intelligence (NL4AI), Rome*.

Chakraborty, S., Bedi, A., Zhu, S., An, B., Manocha, D., & Huang, F. (2024, July). Position: On the possibilities of AI-generated text detection. In *Proceedings of the Forty-first International Conference on Machine Learning*.

Chatterjee, S., & Bhattacharya, K. (2020). Adoption of artificial intelligence in higher education: A quantitative analysis using structural equation modelling. *Education and Information Technologies, 25*, 3443–3463. https://doi.org/10.1007/s10639-020-10152-1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Das, B., & Chakraborty, S. (2018). An improved text sentiment classification model using TF-IDF and next word negation. *arXiv preprint arXiv:1806.06407*.

Dave, R., & Balani, P. (2015, March). Survey paper of different lemmatization approaches. In *International journal of research in advent technology science and technology. Special Issue: First International Conference on Advent Trends in Engineering, Science and Technology, ICATEST* (pp. 366-370).

Delgado-Chaves, F. M., Jennings, M. J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z. M., Baumbach, J., & Baumbach, L. (2025). Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America, 122*(2), e2411962122. https://doi.org/10.1073/pnas.2411962122

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060

Feuerriegel, S., Hartmann, J., Janiesch, C., Buxmann, P., Heinzl, A., Leimeister, J. M., & Niehaves, B. (2024). Generative AI. *Business & Information Systems Engineering, 66*(2), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint* arXiv:2301.07597. https://arxiv.org/abs/2301.07597

Hua, S., Jin, S., & Jiang, S. (2024). The limitations and ethical considerations of ChatGPT. *Data Intelligence, 6*(1), 201–239. https://doi.org/10.1162/dint_a_00243

Hutson, J. (2024). Rethinking plagiarism in the era of generative AI. *Faculty Scholarship, 619*. https://digitalcommons.lindenwood.edu/faculty-research-papers/619

Ide, K., Hawke, P., & Nakayama, T. (2023). Can ChatGPT be considered an author of a medical article? *Journal of Epidemiology, 33*(7), 381–382. https://doi.org/10.2188/jea.JE20230030

International Committee of Medical Journal Editors. (n.d.). *Defining the role of authors and contributors*. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.htmlI

Islam, I., & Islam, M. N. (2024). Exploring the opportunities and challenges of ChatGPT in academia. *Discover Education, 3*, 31. https://doi.org/10.1007/s44217-024-00114-w

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jarrah, A. M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies, 13*(4), e202346. https://doi.org/10.30935/ojcmt/13572

Karwatowski, M., & Pietron, M. (2022). Context based lemmatizer for Polish language. *arXiv preprint arXiv:2207.11565*.

Krzeszewska, U., Poniszewska-Marańda, A., & Ochelska-Mierzejewska, J. (2022). Systematic comparison of vectorization methods in classification context. *Applied Sciences*, *12*(10), 5119.

Mahajan, S. (2023). Artificial intelligence and its impacts on the society. *Contemporary Social Sciences*, 32(4), 135–149. https://jndmeerut.org/wp-content/uploads/2024/01/10-6.pdf

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., Qureshi, F., & Cheungpasitporn, W. (2024). Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: A narrative review. *Clinics and Practice, 14*(1), 89–105. https://doi.org/10.3390/clinpract14010008

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining, 16*, 20. https://doi.org/10.1186/s13040-023-00339-9

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mondal, H., Mondal, S., & Mittal, P. (2024). Evaluating large language models for selection of statistical test for research: A pilot study. *Perspectives in Clinical Research, 15*(4), 178–182. https://doi.org/10.4103/picr.picr_275_23

Nguyen, T., Hatua, A., & Sung, A. (2023). How to detect AI-generated texts? In *2023 IEEE 7th UEMCON* (pp. 0464–0471). https://doi.org/10.1109/UEMCON59035.2023.10316132

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Rani, R., & Lobiyal, D. K. (2022). Performance evaluation of text-mining models with Hindi stopwords lists. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2771-2786.

Rashid, A., & Sodhi, G. K. (2024). Classification of human written text and artificial intelligence generated text using machine learning. *International Journal of Creative Research Thoughts, 12*(7), f632–f638.

Salvagno, M., Taccone, F.S. & Gerli, A.G. Can artificial intelligence help for scientific writing?. *Crit Care* **27**, 75 (2023). https://doi.org/10.1186/s13054-023-04380-2

Shah, A., Ranka, P., Dedhia, U., Prasad, S., Muni, S., & Bhowmick, K. (2023). Detecting and unmasking AI-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications, 14*(10).

Shijaku, R., & Canhasi, E. (2023). *ChatGPT generated text detection*. https://doi.org/10.13140/RG.2.2.21317.52960

Sikander, B., Baker, J. J., Deveci, C. D., Lund, L., & Rosenberg, J. (2023). ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. *Cureus*, *15*(11).

Sivesind, N. T., & Winje, A. B. (2023). Turning poachers into gamekeepers: Detecting machine-generated text in academia using large language models (Bachelor's thesis, Norwegian University of Science and Technology [NTNU]).

Tatzel, A., & Mael, D. (2023). *'Write a paper on AI plagiarism': An analysis on ChatGPT and its impact on academic dishonesty in higher education*. https://www.lasell.edu/documents/Writing%20Program/2023%20Winners/TatzelA%20100%20Level%20Winnter%202023.pdf

Webb, G. I., Keogh, E., Miikkulainen, R., & Sebag, M. (2011). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 713–714). Springer. https://doi.org/10.1007/978-0-387-30164-8_576

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 1–66.

Yildiz Durak, H., Eğin, F., & Onan, A. (2025). A comparison of human-written versus AI-generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education, 60*(1), e70014.