replika

DOI: 10.32564/134.2

2024 (134): 23–34.

replika.hu/replika/134

Katona Eszter, Szeitl Blanka és Túry-Angyal Emese

A kutatás ára

Az empirikus társadalomkutatás ökológiai lábnyoma és ennek mérési lehetőségei a természetesnyelv-feldolgozás példáján keresztül

Absztrakt: Az empirikus társadalomkutatás eszközei (adatgyűjtések és elemzési módszerek) hasonlóan járulnak hozzá a környezeti terheléshez, mint például a közlekedés, a fogyasztás vagy az állattenyésztés. Az utóbbiak kapcsán intenzíven vizsgáljuk ökológiai lábnyomuk mértékét, illetve annak csökkentésének különböző lehetőségét. Ezzel szemben, jelenleg a társadalomtudományi gyakorlatban elsősorban a kutatás költsége, idő- és humánerőforrás igénye határozza meg az eszközök kiválasztását. A felvett adatok mellett hatalmas mennyiségben jön létre tranzakciós vagy éppen social média típusú adat, melyek esetében az elemzések során az összetett modellek számítási kapacitása, illetve a bevont paraméterek száma különböző mértékű környezeti terhet jelenthetnek. A társadalomtudományi kutatások ökológiai lábnyomának meghatározása kulcsfontosságú lépés lehet a fenntartható tudományos gyakorlat kialakításában. Ebben a cikkben áttekintjük a téma kapcsán releváns nemzetközi irodalmat, kiemelve a már meglévő bevált gyakorlatokat. Elemzésünkben a természetesnyelv-feldolgozás mint népszerű és folyamatosan fejlődő elemzési módszer példáján keresztül azt mutatjuk meg, hogy a társadalomtudományi kutatások ökológiai lábnyomának tudatosabb figyelembevétele és csökkentése hozzájárulhat a fenntartható tudományos gyakorlat kialakításához. Elemzésünk segít kialakítani egy olyan szemléletet, ami csökkentheti a kutatások környezeti terhelését, anélkül, hogy veszélyeztetné a tudományos eredmények minőségét és megbízhatóságát.

Kulcsszavak: empirikus társadalomkutatás, ökológiai lábnyom, fenntartható tudomány, természetesnyelv-feldolgozás

Bevezetés

Az empirikus társadalomkutatás eddigi történetében két olyan fordulópontot említhetünk, melyek kifejezetten nagy hatással voltak a módszerekre. Az egyik ilyen a survey-módszerek térnyerése, amit globális szinten a ’60-as évekre datálunk. A másik fordulópont a ’90-es évekre tehető, amikor – sok más folyamattal párhuzamosan – megjelent az organikus (organic) vagy más néven talált (found) adat. Ez ugrásszerű változást okozott az elemzési módszerek fejlődésében is: komplexebb és egyre komolyabb számítási kapacitást igénylő módszerek jelentek meg. Ilyenek például a *machine learning* (Machine Learning, ML), illetve a természetesnyelv-feldolgozás (Natural Language Processing, NLP) technikái. Egyre relevánsabb az, hogy milyen előnyei és hátrányai vannak az új módszereknek, illetve hogy milyen szempontok mentén választható ki a legalkalmasabb elemzési eszköz. Kmetty (2018) dolgozata öt szakmai szempont mentén mutatja be az új módszerek helyét a szociológiában, melyek az időbeliség, a vizsgált populáció vagy jelenség, a mintanagyság, a használt statisztikai mutatók és a kutatási terület. Ezen szakmai szempontok mellett a kutatási módszerek választásánál fontos dimenzió a költség is. A legtöbb, módszereket összehasonlító tanulmány kitér arra, hogy bizonyos adatgyűjtési módszerek egymáshoz viszonyítva milyen költségterhelést jelentenek egy kutatás esetében, és hogy ezek a költségek milyen viszonyban állhatnak az adatok minőségével, megbízhatóságával, illetve felhasználhatóságával. A költséget viszont nem csak pénzben lehet mérni: úgy, ahogy az a közlekedés, a fogyasztás vagy az állattartás esetében is releváns, a kutatások esetében is érdemes megbecsülni vagy legalább átgondolni a kutatás potenciális környezeti terhelését mint költséget, és ezt a dimenziót is bevonni a módszerekkel kapcsolatos döntési szempontok körébe. Világszerte az internet használatának szénlábnyoma 28-tól 63 gramm szén-dioxiddal egyenértékű gáz kibocsátásáig terjed gigabájtonként, míg víz- és területlábnyoma gigabájtonként 0,1-től 35 literig és 0,7-től 20 négyzetcentiméterig terjed (Obringer et al. 2021). Egy egyórás Zoom-hívás 150 és 1000 gramm közötti szén-dioxid-kibocsátást eredményezhet, és a hívás karbonlábnyoma 96%-kal csökkenthető, ha a kamera nincs bekapcsolva. A generatív mesterséges intelligencia, mint például a ChatGPT, használata, 4-5-ször több energiát igényel, mint egy hagyományos webes keresés, egyelőre azonban ez a szempont nem jelenik meg a kutatások tervezésekor. Elemzésünkben ezért összegyűjtjük az eddigi nemzetközi tapasztalatokat a kutatások környezeti terhelése szempontjából, és bemutatjuk mindezt egy kiválasztott módszer, a természetesnyelv-feldolgozás példáján.

Tanulmányunk célja az, hogy rávilágítson egy olyan kérdésre, amelyet egyelőre szervezett szinten nem értékelünk ki egy kutatás esetében, és amely nem jelenik meg a tudományos diskurzusban. A tanulmány a következő szerkezetet követi: Az 1. fejezetben bemutatjuk a természetesnyelv-feldolgozás módszerét és ismertetjük a tanulmányban megjelenő szakkifejezéseket, rövidítéseket. A 2. fejezetben azzal foglalkozunk, hogy milyen szempontokon múlik a természetesnyelv-feldolgozás esetében a kutatások ökológiai lábnyoma, a 3. fejezetben pedig az ökológiai lábnyom mérésének, kvantifikálásának és riportálásának lehetőségeit mutatjuk be. Ezután a 4. fejezetben ismertetjük azokat a módszereket és lehetőségeket,

amikkel a kutatásokat környezeti szempontból fenntarthatóbbá tehetjük, illetve amiket érdemes egy kutatás tervezésénél (és a végrehajtás közben) átgondolni. Az elemzés végén összegezzük a témával kapcsolatos eredményeket és a potenciális jövőbeli irányokat.

A természetesnyelv-feldolgozás módszere

A természetesnyelv-feldolgozás (NLP) fejlődésének történetét Manning (2022) négy nagy korszakra osztja: (1) 1950–1969: a gépi fordítás korszaka; (2) 1970–1992: a kézzel épített, szabályalapú rendszerek korszaka; (3) 1993–2012: a statisztikai alapú, gépi tanulást használó modellek korszaka; (4) 2013–jelen: a deep learning (mélytanulás), neurális hálók, transzformerek korszaka. A legutolsó korszak kétfelé bontható: 2013–2017 között a deep learning modellek, a neurális hálók domináltak, míg a transzformerek korszaka inkább 2018 után kezdődött.

Az NLP kutatások az első korszakban az automatikus beszédfelismerésre és a gépi fordításra (elsősorban más nemzetek tudományos eredményeinek fordítására) fókuszáltak (Manning 2022; Bender et al. 2021). Ekkor a számítási kapacitás és a felhasznált adatmennyiség is elenyésző volt, így ezek a módszerek szótáralapon, valamint egyszerűbb, előre definiált szabályok követésével működtek – figyelmen kívül hagyva a komplexebb nyelvtani szabályokat. A második korszak már ezzel szemben főképp különböző, az emberi nyelvre jellemző, komplexebb jelenségek, mint például a szintaxis (azaz a mondatok szerkezeti felépítésének) megértésére fókuszált – de még hasonló módszerek alkalmazásával, mint az első időszakban. A harmadik korszakban, a kutatások iránya jelentősen megváltozott: nagy mennyiségben váltak hozzáférhetővé a digitális szövegek, és a kutatások fő módszerei az empirikus, gépi tanulás alapú NLP-modellek lettek, amelyek a mai napig uralják ezt a mezőt (Manning 2022). A deep learning és a neurális hálózatok bevezetése alapvetően változtatta meg a munkát. Az időszakra jellemző egyik nagy lépés a szóbeágyazási modellek használata volt. A módszer tulajdonképpen egy dimenziócsökkentő eljárás, melynek segítségével egy kifejezés kontextusát ragadhatjuk meg (Kmetty 2022). Ebben a megközelítésben a szavakat és mondatokat több száz vagy ezer dimenziós, valós számokból álló vektortérben helyezik el, és a jelentésbeli vagy szintaktikai hasonlóságokat a térbeli közelség/távolság reprezentálja (Manning 2022).

2018-ban minden megváltozott a neurális háló alapú transzformer-modellek megjelenésével és a nagy nyelvi modellek (Large Language Models, LLM) elterjedésével. A transzformer kifejezés a modell mögöttes neurális hálózati struktúrájára utal. A transzformer egy konkrét mélytanulási architektúra, amelyet Vaswani és társai 2017-ben mutattak be a „Attention is All You Need" című cikkben. Az LLM-ek nagyon nagy méretű modellek, amelyeket óriási korpuszokon tanítottak be, hogy megértsék és generálni tudják az emberi nyelvet. A legtöbb LLM a transzformer-architektúrát használja, de a transzformereket más célokra is fel lehet használni, pl. képfelismeréssel kapcsolatos feladatokra. Ilyen modell például a BERT (Bidirectional Encoder Representations from Transformers) és a GPT (Generative Pre-Trained Transformers).

Az ökológiai lábnyom mérése a természetesnyelv-feldolgozás területén

Az ökológiai lábnyom mérése a természetesnyelv-feldolgozás területén azért kiemelten fontos, mert az LLM módszerek megjelenése jelentős változást hozott a nyelvi modellek teljesítménye tekintetében, mivel a deep learning kutatásokhoz szükséges számítások 300 000-szeresükre nőttek 2012-ről 2018-ra (Schwartz et al. 2019). Az egyre többet emlegetett OpenAI, GPT-3 modellek például több száz milliárd, illetve trilliós nagyságrendű paraméterrel rendelkeznek, ennek köszönhetően a teljesítményük rendkívül magas – de ezzel nagyobb méret és nagyobb energiafogyasztás is párosul (Manning 2022; Bender et al. 2021). Az 1. táblázat a legismertebb modellcsaládok bemutatására fókuszál a megjelenés éve, a paraméterek száma és az adat mérete mentén.

1. táblázat. Példák a nagy nyelvi modellekre: paraméterek, adatméret, és tanító korpusz alapján

Év	Modell	Paraméterek száma[1]	Adat mérete	Tanító korpusz
2018	GPT-1	120 millió	–	BooksCorpus
2019	GPT-2	1,5 milliárd	–	Reddit
2019	BERT	340 millió	16 GB	BooksCorpus, Wikipedia (angol)
2019	ALBERT	223 millió	16 GB	BooksCorpus, Wikipedia (angol)
2019	RoBERTa	355 millió	161 GB	BooksCorpus, Wikipedia (angol), CC-NEWS, STORIES, Reddit
2020	GPT-3	175 milliárd	570 GB	Common Crawl (szűrt), WebText2, Books1, Books2, Wikipedia
2021	Switch-C	1,57 billió	745 GB	–
2022	PaLM	540 milliárd	780 GB	Dokumentumok a webről, könyvek, Wikipedia, „conversational data”,[2] GitHub kódok
2022	BLOOM	176 milliárd	350 GB	ROOTS
2023	LLaMA 1	65 milliárd	–	Online források
2023	LLaMA 2	70 milliárd	2 TB	Online források
2023	GPT-4	1,76 billió	–	–
2024	PaLM 2	340 milliárd	1,5 TB	Dokumentumok a webről, könyvek, kódok, matematikai adatok, „conversational data”
2024	LLaMA 3	405 milliárd	2 TB	Online források

Forrás: saját szerkesztés Bender et al. 2021; Minaee et al. 2024; Naveed et al. 2024 és saját adatgyűjtés alapján

1 Minden modelltípusból csak a legnagyobb elérhető paraméterszámot adtuk meg.

2 A „conversational data” alatt olyan adatokat értünk, amelyeket digitális beszélgetésekből (e-mailekből, chat beszélgetésekből, hívásokból) nyertek ki.

A táblázatban látható, transzformer alapú, (általában többnyelvű) modellek mind fontos szerepet játszottak a nagy nyelvi modellek fejlődésében, hiszen az új modellek új módszereket, architektúrákat jelentenek. A modellek között vannak nyílt forráskódú (pl. LLaMA család) és zárt forráskódú modellek (pl. a GTP család) egyaránt. Ezzel láthatóvá válik egyrészt a (tudományos) fejlesztői közösség, másrészt a nagy cégek szerepe is az LLM-ek fejlődésében. A BERT, a GPT és a PaLM modellsorozatok például úttörő megoldásokat hoztak a nyelv megértésében és szövegek generálásában egyaránt.

A fenti modellek jellemzőiből kiolvasható az LLM-ek evolúciója, melynek egy indikátora például a paraméterek számának növekedése. A paraméterek a modellek változó értékei, amelyek azt befolyásolják, hogyan értelmezi a modell az adatokat. Ahogy a modellek tanulnak, úgy változnak ezek az értékek, hogy a lehető legpontosabb eredményeket érje el a modell. Így tehát minél több paraméter van egy modellben, annál összetettebb összefüggéseket képes megtanulni és annál pontosabban működik. Az első modellekhez képest (mint például az első, 2018-as GTP modell 120 millió paraméterrel) az újabb modellek számottevően több paraméterrel rendelkeznek[3] ami egyrészt jelentős előrelépést jelent a modellek teljesítményében, de a környezeti terhelésében is.

A fenti modellek nagyon sok és sokféle szövegen lettek betanítva: az online szövegektől, beszélgetésektől kezdve a Wikipédián át egészen a kódokig. Az adatméret növekedése szintén jelzi azt a trendet, hogy minél több információt integráljanak a modellekbe a jobb teljesítmény érdekében.

Látható, hogy az LLM modellek jelentős számítási kapacitást igényelnek, és ahogy a modellek egyre nagyobbak és összetettebbek lesznek, a tanításhoz és működtetéshez szükséges energiafogyasztás is egyre növekszik. Az LLM tanításához nem elég egy irodai laptop, ehhez nagyobb teljesítményű számítógépre vagy szerverre van szükség, ami akár heteken át is futhat. Mivel a nyelvi modellek mérete és számítási igénye folyamatosan növekszik, az NLP környezeti lábnyoma jelentős kérdéssé vált az utóbbi években (Bender et al. 2021).

Az LLM-ek energiafelhasználását több tényező befolyásolja (Stojkovic et al. 2024; Patterson et al. 2021; Strubell, Ganesh és McCallum 2019):

1. *Adatközpontok hatékonysága:* Az adatközpontokra az LLM-ek kontextusában azért van szükség, mert ezek biztosítják a nagy teljesítményű számítástechnikai infrastruktúrákat (számítási és pl. korpusztárolási kapacitás). Az energiafelhasználás szempontjából fontos az energiaforrás típusa (megújuló vagy nem megújuló) és a központ hatékonysága is. Az adatközpontok olyan processzorokkal és hardverekkel vannak felszerelve, amelyek lehetővé teszik ezeknek a számítási feladatoknak az elvégzését.
2. *Hardver hatékonysága:* A különböző processzorok és célhardverek eltérő energiahatékonyságúak, valamint ezeknek a hűtéséhez is energia szükséges. Az utóbbi időben a kevésbé integrált CPU (Central Processing Unit) szerepét egyre inkább átveszi a GPU (Graphics Processing Unit), ami sokkal hatékonyabban használható a különböző modellek tanítására, mivel a számításokat végző egység és az eredményeket tároló nagy sebességű memória egy rendkívül gyors kapcsolaton keresztül képes kommunikálni.

3 Például a 2023-as megjelenésű GTP-4 1,76 billió paraméterrel rendelkezik, tehát nagyjából 15 ezerszer több paraméterrel, mint az első modell.

A modell tanításához GPU vagy TPU (Tensor Processing Units) használata javasolt. Az LLM-ek nagy számítási igénye hatékonyan kezelhető GPU-k segítségével, mert ezek alkalmasak párhuzamosított számítások végrehajtására. A GPU-k nagy számítási teljesítményüknek köszönhetően jelentősen felgyorsítják a tanítási folyamatot. A CPU-nak (Central Processing Unitnak) az LLM-ek esetében általában az adatfeldolgozás során van hangsúlyosabb szerepe, de az LLM modellek kiértékelése és finomhangolása gyakran CPU-n történik (Youvan 2023).

3. *Modellméret:* A nagyobb paraméterszámú modellek több számítást igényelnek, így több energiát is fogyasztanak, hiszen több adatot használnak a tanuláshoz és finomhangoláshoz egyaránt. Ez egyrészt több számítási kapacitást, másrészt nagyobb memóriát, hosszabb futásidőt és több energiát igényel. Emellett a nagyobb adathalmazok tárolása is növeli az energiafelhasználást. A nagyobb modellek nemcsak a tanítás során használnak több energiát, de az inferencia során is, tehát amikor a modellt futtatják, használják.
4. *Finomhangolás:* Minél nagyobb egy modell, annál több iterációra és finomhangolásra van szükség az optimális teljesítmény elérése érdekében. Mivel minden iteráció során a teljes modell összes paramétere frissül, minden egyes iteráció jelentős futásidőt, azaz energiahasználatot jelent. A modell gyakori újratanítása további pluszenergiát igényel.
5. *Felhasználás (inferencia):* A modell minden egyes lekérdezésével energiát használunk, így tehát érdemes megfontolni, hogy valóban a ChatGPT-re van-e szükségünk, vagy egy Google-kereséssel vagy a polcról levett receptkönyvvel is megoldható-e az esti vacsora megtervezése.

Az ökológiai lábnyom mérésének, kvantifikálásának és riportálásának lehetőségei

Az ökológiai lábnyom valamilyen tevékenység teljes környezeti terhelését jelenti, ami magába foglalja az energiafogyasztás, a szén-dioxid-kibocsátás és a karbonlábnyom mérését. Ennek kvantifikálása komplex feladat, melyre többféle módszer létezik. Az energiahasználat Joule-ban vagy wattórában mérhető, és az összes felhasznált energia mennyiségét mutatja meg egy adott feladat, például egy modell tanítása során (Henderson et al. 2020). A szén-dioxid-kibocsátás konkrét CO_2-mennyiségeket jelöl, míg a karbonlábnyom többféle különböző gázt, például metánt és nitrogén-oxidokat egyaránt vizsgál. A karbonlábnyomot ezért CO_2-egyenértékben (CO_2-eq) fejezzük ki. A szén-dioxid-ekvivalens vagy CO_2-ekvivalens, rövidítve CO_2-eq, egy olyan metrikus mérőszám, amelyet különböző üvegházhatású gázok kibocsátásának összehasonlítására használnak, azok globális felmelegedési potenciálja alapján, más gázok mennyiségét átváltva az azonos globális felmelegedési potenciállal rendelkező szén-dioxid egyenértékű mennyiségére.[4] A karbonlábnyom azt méri, hogy az adott tevékenység milyen mértékben járul hozzá az üvegházhatású gázok kibocsátásához és így a klímaváltozáshoz. Vizsgálata azért fontos,

4 https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Carbon_dioxide_equivalent (letöltve: 2025. január 14.).

mert széleskörűen használt fogalom a termékek és szolgáltatások életciklusához kapcsolódó szén-dioxid-egyenérték-kibocsátások és -hatások jelentésére.

Az internethasználat elterjedésével együtt az adatforgalom sebessége és az ezt biztosító adatközpontok (data centerek) mérete is nőtt (Ristic, Madani és Makuch 2015). Az egyre szélesebb körben használt nagy nyelvi modellek energiaigénye, szén-dioxid-kibocsátása és karbonlábnyoma egyre jobban kutatott terület, mérésükre több módszert is fejlesztettek már a kutatók. Ezeknek egyik alapja az energiaigény mérése, ami azért is fontos, mert jól dokumentált terület, és abból lehet következtetni a szén-dioxid-kibocsátás mértékére.

Strubell, Ganesh és McCallum (2019) esettanulmányukban a legnépszerűbb NLP modellek energiafogyasztását vizsgálták a modellek fejlesztéséhez szükséges összes energiafogyasztás segítségével. Az energiafogyasztás mérése során a modelleket alapértelmezett beállításokkal tanították, miközben mintavételeken keresztül figyelték a GPU és CPU teljesítményfogyasztását, majd a megfigyeléseiket átlagolták. A kutatás során a modellek teljes tanításhoz szükséges időt a modellt bemutató, eredeti cikkekben megadott tréningidők és a használt hardver alapján becsülték meg. A U.S. Environmental Protection Agency által megadott adatok alapján az energiafogyasztásból származó CO_2-kibocsátást (font/kilowattóra) becsülték meg. Ez az átváltás figyelembe veszi az Egyesült Államokban az energia előállításához használt különböző energiaforrások (elsősorban földgáz, szén, nukleáris és megújuló) arányát. Az energiafogyasztást kilowattórában (kWh) úgy becsülték, hogy összeadták a GPU-, CPU- és DRAM-[5] fogyasztást, majd ezt megszorozták a Power Usage Effectiveness (PUE) együtthatóval, amely figyelembe veszi a számítástechnikai infrastruktúra támogatásához (főként hűtéshez) szükséges további energiát.

Henderson és szerzőtársai (2020) a szén-dioxid-kibocsátás kiszámításához szintén a felhasznált energia mennyiségéből indulnak ki a kilowattórában mért teljesítményt megszorozzák a helyi energiahálózat karbonintenzitásával.A karbonintenzitás[6] (carbon intensity) azt mutatja, hogy mennyi szén-dioxid vagy más üvegházhatású gáz kibocsátása történik egy adott mennyiségű energia előállítása során, általában gramm CO_2-ekvivalens/kilowattóra (gCO_2-eq/kWh) egységben fejezik ki.

Az energiahálózatok szénintenzitásának mérésére egy nyílt forráskódú térképet[7] használnak, ami alapján meghatározzák a régiókat, és a legkisebb behatárolható régiót használva becsülhető meg egy adott helyszín szénintenzitása. Például, ha egy kísérletet San Franciscóban futtatnak, és elérhető az átlagos karbonintenzitás mind az USA-ra, mind Kaliforniára vonatkozóan, akkor az utóbbit használják. Az energiahálózatok között jelentős különbségek vannak a széndioxid kibocsátás tekintetében. CO_2-eq-intenzitás tekintetében azonnali csökkenést lehet elérni a szénkibocsátásban, ha minden tréningmunkát karbonhatékony energiahálózatokra helyeznek át. A 2017-es átlagok alapján például egy munkafolyamat Quebecben történő futtatása 30-szor alacsonyabb szénkibocsátást eredményezne, mint ugyanez Észtországban (Henderson et al. 2020).

---

5 A DRAM (Dynamic Random-access Memory) az egyik leggyakrabban használt memóriatípus. Dinamikusan tárolja az adatokat, így azokat gyorsan elérhetővé tudja tenni az LLM-ek működéséhez. A DRAM energiafelhasználása kisebb, mint sok merevlemezé, azonban a frissítési ciklusok miatt folyamatos energiaellátást igényel.

6 https://carbonintensity.org.uk/ (letöltve: 2025. január 14.).

7 Ez a honlap itt érhető el: https://www.electricitymap.org (letöltve: 2025. január 14.).

Henderson és szerzőtársai hangsúlyozzák, hogy a szén-dioxid-kibocsátás követhetőbbé válásához a felhasznált energia dokumentálása az első lépés. Ezt négy szempontból különböztetik meg, egyrészt a környezeti terhelésre következtethetünk a felhasznált energia mennyiségéből (Joule-ban vagy Wattban mérve); a használt számítási kapacitásból (floating point operations per second, FPOs, GPU-/CPU-használat százalékos arányban; GPU-órákban megadva); a futási időből (runtime); illetve a szén-dioxid-kibocsátásból. Ezek a mérőszámok kombinálhatók is, de különböző eredményekre vezetnek, ha összehasonlítjuk őket. Ezért Henderson és szerzőtársai (2020) az experiment-impact-tracker nevű frameworkkel a környezeti terhelés mérését hivatottak konzisztenssé, egységessé tenni, megkönnyíteni. Módszerükben egy projekt karbonmérlege (carbon accounting) a szén-dioxid és egyéb üvegházhatású gázok kibocsátásának és a projekt karbon-ellentételezésének (carbon offsetting) különbsége. Ezt CO_2-eq mértékegységben adják meg. Ez az a szénmennyiség – és más üvegházhatású gázok átváltott mennyisége –, amely a projekt eredményeként kerül a légkörbe. A karbonmérleg ennek és a projekt által offsetelt/kiváltott, megtakarított kibocsátás mennyiségének különbségét jelenti. Például egy vállalat megvásárolhatja a projektjéhez a szükségesnél több megújuló energiát, hogy ellentételezze a hozzájárulását a szénkibocsátáshoz. A frameworkben továbbá a szén-dioxid-kibocsátás társadalmi költségeit is mérik, ehhez az Egyesült Államok Környezetvédelmi Ügynöksége (EPA, 2013) által is használt SC-CO_2 mutatót veszik figyelembe. Ez az egy adott évben kibocsátott tonna szén-dioxid által okozott hosszú távú károk és a CO_2-csökkentés előnyeit is tartalmazza.

A kutatók számára Lacoste és szerzőtársai (2019) létrehoztak egy weboldalt a gépi tanulás során használt modellek karbonlábnyomának mérésére, ami a hardver típusa, a futási idő (órában), a szolgáltató (pl. Google Cloud, saját infrastruktúra), valamint a régió alapján kiszámítja a modell karbonlábnyomát. A modellükben két szempont alapján értékelik ki a szén-dioxid (vagy azzal ekvivalens hatású) gázok kibocsátását: egyrészt az energia típusát, másrészt a számítási környezetet és tréningidőt veszi figyelembe. A honlapon standardizált jelentést mint egy hivatkozásgenerátort is létre lehet hozni, amelyet a cikk/blogposzt végére javasolnak írni a szerzők. A korábbi kutatásokkal egybehangzóan azt állítják, hogy a kibocsátás mérése az első lépés a tudatos felhasználáshoz. A standardizált jelentéskészítés minden cikk esetében elengedhetetlen, hogy tartalmazzon szabványosított jelentést az energia- és szénkibocsátásról, beleértve a kísérletekből származó összes kibocsátást. Ez reálisabb képet ad a felhasznált erőforrásokról (Henderson et al. 2020).

Fenntarthatóságért tehető lépések a kutatásokban

A kutatások zöldítése érdekében számos lehetőség áll rendelkezésre, amelyek túlmutatnak a felhasznált energia közlésén. Lacoste és szerzőtársai (2019) négy további pontot említenek, amelyek között szerepel a kibocsátás kvantifikálása, a felhőszolgáltatók és adatközpontok helyének gondos kiválasztása, a források tudatos felhasználása (például grid search csak indokolt esetben), valamint a hatékony hardverek használata. A kibocsátás mérése a korábbiakban említett szempontok miatt fontos: akkor lehet tudatosan modellt és számítógépes kapacitást választani, ha tudjuk, milyen energiafelhasználással, illetve

karbonlábnyommal dolgozunk. A felhőszolgáltatók és adatközpontok helyének megválasztása szintén szempont lehet a kutatás tervezésekor: az energiahatékonyság nagyban függ attól, hogy az adott országban, ahol a szerver található, mi a legfőbb energiaforrás – itt is lehet törekedni a megújuló energia választására. Strubell, Ganesh és McCallum (2019) három pontban foglalják össze a legfontosabb javaslataikat: (1) az NLP gépi tanulási modelleknél jelenteni kell a tanulási időt és a modell érzékenységét a hiperparaméterekre; (2) az akadémiai kutatóknak egyenlő hozzáférést kell biztosítani a számítási erőforrásokhoz; (3) a kutatóknak prioritásként kell kezelniük a hatékony modellek és hardverek fejlesztését. A három ajánlás összefügg, hiszen a kutatók csak akkor tudnak felelős döntést hozni a számítás módjáról, ha van alternatívájuk, és hozzáférnek az adatok mellett a hatékony eszközökhöz is. Dodge és szerzőtársai (2022) szintén amellett érvelnek, hogy ha a kutatók hozzáférnének a számítási tevékenységeik CO_2-kibocsátásáról szóló információkhoz, döntéseiket úgy tudnák alakítani, hogy csökkentsék ezeket a kibocsátásokat, anélkül hogy a feladatok elvégzéséhez szükséges számítást csökkentenék.

Henderson és szerzőtársai (2020) az energiahatékonyságot ösztönző rendszerszintű változásokat javasolnak a gépi tanulási rendszerekben. Ezek között szerepel az energiatakarékos beállítások alapértelmezetté tétele a gyakran használt platformokon és modellekben, valamint a teljesítménynövekedésből származó nyereség szén-dioxid-kibocsátás tekintetében történő költségeinek kiszámítása is. A kutatók számára fontos lehet egy hatékony tesztelési környezet kialakítása, illetve reprodukálható kutatások végrehajtása: a kódok és modellek közzététele segít elkerülni a további szén-dioxid-kibocsátást. Az eredmények reprodukálása a kutatás fontos része, és a közzététel hiánya miatt több energiát kell felhasználni, különösen nagy modellek esetében.

Az energiafelhasználásról szóló jelentések publikálása is ösztönzőleg hathat, amelyek alapján Henderson és szerzőtársai (2020) javasolják az éghajlatbarát kutatási jelvények bevezetését a konferenciákon, ezáltal is elismerve azokat a cikkeket, amelyek jelentős erőfeszítéseket tesznek hatásuk mérséklésére. Például egy számításigényes cikk, ha bizonyítja, hogy erőforrásait tiszta régiókban futtatja, ilyen jelvénnyel jutalmazható. Más jelvényeket olyan cikkeknek lehet adni, amelyek energiatakarékos algoritmusokat fejlesztenek, amelyek hasonló teljesítményt nyújtanak, mint a legkorszerűbb megoldások.

Konklúzió

Tanulmányunkban bemutattuk, hogy a digitalizáció hatására a számítógépes adatfeldolgozás egyre elterjedtebbé vált a kutatásokban, hasonlóan az élet más területeihez. Míg más területeken már egyre inkább a figyelem középpontjába került ennek környezeti terhelése, a társadalomtudományos kutatások körében ez még kevéssé kutatott téma. Azonban a kutatásokban használt számítógépes nyelvi modellek esetében már léteznek számítások a környezeti terhelésről. Ezen a példán keresztül mutattuk be, hogy milyen lépéseket lehet tenni a tudatosabb és fenntarthatóbb kutatások érdekében: több tanulmány szorgalmazza a kutatások energiaigényének kutatók által történő jelentését, hiszen ebből már létező kalkulátorok segítségével a környezeti terhelés szempontjából fontos karbonlábnyomot ki lehet számolni, és így a kutatás tervezésekor lehet számolni vele, optimalizálni.

A számításokat végző adatközpontok, szerverparkok helyének figyelembevételével a felhasznált energia típusa is követhető, valamint további konkrét javaslatok is születtek már az energiafelhasználás tudatosítására, amelyek más területeken már bevett szokásnak számítanak, viszont a társadalomtudományi kutatásokból egyelőre hiányoznak.

A társadalomkutatók (pl. Németh 2015) sokszor hangsúlyozzák, hogy a szakterület sajátosságai miatt a modellek értelmezhetősége fontosabb, mint önmagában egy modell pontossága. Az utóbbi években így a komplex, gépi tanulás során alkalmazott modelleknél ez a szempont sokat fejlődött. A fent bemutatott esettanulmányok tanulságai mentén jó lenne, ha a fenntarthatóság szempontja – ami természetesen nem csak a szociológiai kutatások szempontjából fontos – a következő években szintén előtérbe kerülne.

Hivatkozott irodalom

Arnfalk, Peter, Ulf Pilerot, Peter Schillander és Peter Grönvall (2016): Green IT in practice: virtual meetings in Swedish public agencies. *Journal of Cleaner Production* 123: 101–112. DOI: https://doi.org/10.1016/j.jclepro.2015.08

Aslan, Josh, Kevin Mayers, Jonathan G. Koomey és Caroline France (2018): Electricity intensity of internet data transmission: untangling the estimates. *Journal of Industrial Ecology* 22(4): 785–798. DOI: https://doi.org/10.1111/jiec.12630

Bender, Emily, Timnit Gebru, Angelina McMillan-Major és Shmargaret Shmitchell (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Pro-ceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,FAccT '21, pages 610–623, New York, NY, USA, March 2021. Association for Com-puting Machinery. ISBN 978-1-4503-8309-7. DOI:10.1145/3442188.3445922. URLhttps://dl.acm.org/doi/10.1145/3442188.3445922

Bieser, Jan C. T., Ralph Hintemann, Lorenz M. Hilty és Severin Beucker (2023): A review of assessments of the greenhouse gas footprint and abatement potential of information and communication technology. *Environmental Impact Assessment Review* 99. DOI: https://doi.org/10.1016/j.eiar.2022.107033

Clément, Léo-Paul Perchard-Villette, Quentin Etienne Samuel Jacquemotte és Lorenz Martin Hilty (2020): Sources of variation in life cycle assessments of smartphones and tablet computers. *Environmental Impact Assessment Review* 84: 106416. DOI: https://doi.org/10.1016/j.eiar.2020.106416

Cabrera-Álvarez, Pablo (2022): Survey Research in Times of Big Data. *Empiria. Revista de metodología de ciencias sociales*. DOI: 10.5944/empiria.53.2022.32611

DOMO (2019): Data never sleeps. Interneten: https://www.domo.com/learn/data-never-sleeps-6 (letöltve: 2025. január 14.).

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario és Will Buchanan (2022): Measuring the Carbon Intensity of AI in Cloud Instances. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT, 22)*. Association for Computing Machinery, New York, NY, USA, 1877–1894. DOI: https://doi.org/10.1145/3531146.3533234

Groves, Robert M. (2011): Three Eras of Survey Research. *Public Opinion Quarterly* 75(5): 861–871. DOI: https://doi.org/10.1093/poq/nfr057

Henderson, Peter, Jin Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky és Joelle Pineau (2020): Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21(248): 1–43. Interneten: https://jmlr.org/papers/volume21/20-312/20-312.pdf (letöltve: 2025. január 14.).

IPCC (2018): Annex I: Glossary [Matthews, John B. R. (szerk.)]. In Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. Masson-Delmotte, Valérie, P. Zhai, Hans-Otto Pörtner, Debra Roberts, Jim Skea, Priyadarshi R. Shukla, Anna Pirani, Wilfran Moufouma-Okia, Clotilde Péan, Robin Pidcock, Sarah Connors, John B.R. Matthews, Youba Sokona, Xingkai Zhou, Mark I. Gomis, Elisabeth Lonnoy, Tom Maycock, Melinda Tignor és Tim Waterfield (szerk.). Cambridge, UK: Cambridge University Press, 541–562. DOI: https://doi.org/10.1017/9781009157940.008

Kmetty Zoltán (2012): A telefonos kutatások speciális problémái. *Statisztikai Szemle: a Magyar Központi Statisztikai Hivatal folyóirata* 90: 41.

Kmetty Zoltán (2022): Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása. *Statisztikai Szemle*100(2): 105–136. DOI: https://doi.org/10.20311/stat2022.2.hu0105

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil és Abe Usher (2015): Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly* 79(4): 839–880. DOI: https://doi.org/10.1093/poq/nfv039

Lacoste, Alexandre, Alexandra Sasha Luccioni, Victor Schmidt és Thomas Dandres (2019): Quantifying the Carbon Emissions of Machine Learning. DOI: https://doi.org/10.48550/arXiv.1910.09700

Malmodin, Jens és Dag Lundén (2018): The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015. *Sustainability.* DOI: https://doi.org/10.3390/su10093027

Minaee, Shervin, Tomáš Mikolov, Narjes Nikzad, Meysam Asgari Chenaghlu, Richard Socher, Xavier Amatriain és Jianfeng Gao (2024): *Large Language Models: A Survey*. Interneten: https://arxiv.org/pdf/2402.06196 (letöltve: 2025. január 14.).

Moberg, Åsa, Christina Borggren és Göran Finnveden (2011): Books from an environmental perspective—part 2: E-books as an alternative to paper books. *International Journal of Life Cycle Assessment* 16(3): 238–246. DOI: https://doi.org/10.1007/s11367-011-0255-0

Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes és Ajmal S. Mian (2023): *A Comprehensive Overview of Large Language Models.* Interneten: https://arxiv.org/pdf/2307.06435 (letöltve: 2025. január 14.).

Némedi Dénes (1999): Empirikus társadalomkutatás a tizenkilencedik század végéig. In *A szociológia kialakulása: tanulmányok.* Felkai Gábor (szerk.). Budapest: Új Mandátum, 429–454.

Németh Renáta (2015): A számok tényleg magukért beszélnek? Hozzászólás Dessewffy Tibor és Láng László írásához. *Replika* 92–93(3–4): 203–208. Interneten: https://www.replika.hu/replika/92-93-15 (letöltve: 2025. január 14.).

Obringer, Renee, Benjamin Rachunok, Debora Maia-Silva, Maryam Arbabzadeh, Roshanak Nateghi és Kaveh Madani (2021): The overlooked environmental footprint of increasing Internet use. *Resources, Conservation and Recycling* 167: 105389. DOI: https://doi.org/10.1016/j.resconrec.2020.105389

Patterson, David A., Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluís-Miquel Munguía, Daniel Rothchild, David R. So, Maud Texier és Jeff Dean (2021): Carbon Emissions and Large Neural Network Training. *arXiv preprint*, arXiv:2104.10350. Interneten: https://api.semanticscholar.org/CorpusID:233324338 (letöltve: 2025. január 14.).

Ristic, Bora, Kaveh Madaniés Zen Makuch (2015): The Water Footprint of Data Centers. *Sustainability.* DOI: https://doi.org/10.3390/su70811260.

Schwartz, Roy, Jesse Dodge, Noah Smith és Oren Etzioni (2020): Green AI. *Communications of the ACM* 63: 54–63. DOI: https://doi.org/10.1145/3381831

Stefkovics Ádám, Szabari Veronika, Batiz Katalin, Bukovics Bence, Pavalacs Anett, Tariska Anett és Zenovitz László (2023): Szociológiai tudástermelés Magyarországon 1971 és 2021 között: Hat társadalomtudományi folyóirat szisztematikus áttekintése. *Szociológiai Szemle* 33(2): 4–28. DOI: https://doi.org/10.51624/SzocSzemle.2023.2.1

Stojkovic, Jovan, Esha Choukse, Chaojie Zhang, Inigo Goiri és Josep Torrellas (2024): Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. *arXiv preprint,* arXiv:2403.20306. Interneten: https://arxiv.org/abs/2403.20306 (letöltve: 2025. január 14.).

Strubell, Emma, Ananya Ganesh és Andrew McCallum (2019): Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint*, arXiv:1906.02243. Interneten: https://arxiv.org/abs/1906.02243 (letöltve: 2025. január 14.).

Suski, Patrick, Jan Pohl és Verena Frick (2020): All you can stream: Investigating the role of user behavior for greenhouse gas intensity of video streaming. *Proceedings of the 7th International Conference on ICT for Sustainability:* 128–138. DOI: https://doi.org/10.48550/arXiv.2006.11129

Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser és Illia Polosukhin (2017): Attention is All you Need. *Neural Information Processing Systems.* Interneten: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (letöltve: 2025. január 14.).

World Economic Forum (2016): *What are the top global risks for 2016?* Interneten: http://www.weforum.org/agenda/2016/01/what-are-the-top-global-risks-for-2016 (letöltve: 2024. június 29.).

Youvan, Douglas (2023): Parallel Precision: The Role of GPUs in the Acceleration of Artificial Intelligence. DOI: https://doi.org/10.13140/RG.2.2.21937.76641

Katona Eszter Rita

statisztikus, az ELTE Társadalomtudományi Karának adjunktusa és kutatója a Research Center for Computational Social Science és a Survey Methods Room Budapest kutatócsoportokban, valamint kvantitatív elemző a Government Transparency Institute-nál.

Szeitl Blanka

statisztikus, az ELTE Társadalomtudományi Kar Statisztika Tanszékének és a Szegedi Tudományegyetem Bolyai Intézetének oktatója, a Survey Methods Room Budapest kutatócsoport kutatója, valamint a HUN-REN Szociológia Intézet tudományos munkatársa.

Túry-Angyal Emese

szociológus, az ELTE Társadalomtudományi Kar Statisztika Tanszékének adjunktusa és a Survey Methods Room Budapest kutatója.