HuBERTUSz: Alacsony paraméterszámú transzformer modellek létrehozása és kiértékelése magyar nyelvre

Ficsor Tamás, Berend Gábor

Szegedi Tudományegyetem, Informatikai Intézet
{ficsort,berendg}@inf.u-szeged.hu

Kivonat Hazánkban is megnőtt az érdeklődés a transzformer modellek alkalmazása iránt. Éppen ezért a modern előtanítási standardoknak megfelelően (pl. dinamikus maszkolás, mondatsorrend-predikció használata) közzétesszük az általunk előállított *tiny*, *small* és *medium* transzformer-variánsokat, amelyeket alapos kiértékelésnek vetettünk alá különböző feladatokon. Eredményeinkből kitűnik, hogy elsődlegesen a tokenosztályozási feladatokon a jóval kisebb paraméterszámmal bíró modellvariánsaink is képesek megközelíteni a nagyságrendekkel több paraméterrel bíró társaik eredményeit.
Kulcsszavak: Kompakt modellek, Előtanítás, Névelemfelismerés, Szófaji egyértelműsítés, Szentimentpredikció, Természetesnyelvi következtetés

1. Bevezetés

A transzformerek (Vaswani és mtsai, 2017) megjelenése nagy hatással volt a természetesnyelv-feldolgozásának folyamatára, amelyek kimagasló teljesítményt képesek elérni, azonban a paraméterek hatékony fölhasználásától még távol állnak. Az hogy pontosan mennyi paraméterrel oldható meg kellő minőségben egy bizonyos feladat, erősen függ annak komplexitásától, továbbá a feladattal kapcsolatban elérhető adat mennyiségétől és minőségétől is (McCoy és mtsai, 2019; Bhargava és mtsai, 2021).

A nyelvi modellek tipikusan több különböző méretkonfigurációban is elérhetők a felhasználók számára (Turc és mtsai, 2019). Egy adott konfiguráció a modell paramétereinek mennyiségét határozza meg, amely többek között a rétegek számát, fejek számosságát és a modellben használt vektorok dimenzióját foglalja magában. Egy-egy probléma megoldására igyekezhetünk az erőforrásaink és a szükségleteink szerint választani ezeket a paramétereket és egy egyedi konfigurációt kapunk. Az egyik legelterjedtebb konfiguráció a magyarra is elérhető ún. alap (*base*), amely ~ 110 millió paramétert tartalmaz.

Sok esetben *„ágyúval lövünk verébre”*, amikor a megoldani kívánt feladatot az elégségesnél – adott esetben nagyságrendekkel – nagyobb paraméterszámmal rendelkező modellel próbáljuk megoldani. Azonban ha rendelkezésre állnak kisebb paraméterszámú modellalternatívák, érdemes lehet először azok használatát feltérképezni. Amennyiben egy kisebb modell használatával is kielégítő eredményt

tudunk elérni, akkor az hosszútávon időbeli, anyagi és környezeti terhelést (CO2 emisszió) csökkentő megtérülésekkel jár.

Ebben a cikkben több, kisebb paraméterszámmal rendelkező magyar nyelvű BERT variáns előtanításával és változatos alkalmazásokban történő kiértékelésével foglalkozunk. Létrehozott modelljeinknek a *Hungarian BERT from University of Szeged* nevet adtuk, amit `HuBERTUSz`-ként rövidítünk. Az előtanítás folyamatát igyekeztünk a modern standardoknak megfelelően előkészíteni, létrehozott modelljeinket pedig különböző, diverz adathalmazokon nyújtott eredmények mentén vetjük össze több magyar nyelvet támogató és könnyen hozzáférhető transzformer modell használatával. Az általunk előtanított modelleket könnyen hozzáférhető módon, a HuggingFace felületén elérhetővé is tesszük[1].

2. Kapcsolódó irodalom

A transzformerek (Vaswani és mtsai, 2017; Devlin és mtsai, 2019) megjelenésével egy új hullám kezdődött számos kutatási területen: mostanra bevett gyakorlattá vált a transzformer-alapú modellek használata például kép-, hang-, és kódfeldolgozási (Dosovitskiy és mtsai, 2020; Hsu és mtsai, 2021; Feng és mtsai, 2020) feladatokon egyaránt. Ezen modellek előállítása rengeteg adatot és számítási erőforrást igényel, ezért a modelleket egyszer előtanítják, és ezt követően csak a kívánt feladatokon finomhangolják őket.

A magyar nyelvvel előtanított egy- és több nyelvet is támogató transzformer modellek is elkezdtek megjelenni az elmúlt évek folyamán. Az egynyelvű, kizárólag a magyar nyelv feldolgozására létrehozott maszkolt nyelvi modellezést végző transzformereket tekintve is számos lehetőség érhető el mostanra (Nemeskey, 2020, 2021; Feldmann és mtsai, 2021; Yang, 2022a; Yang és Váradi, 2021; Ficsor és mtsai, 2022; Orosz és mtsai, 2022; Yang és mtsai, 2022; Yang, 2022b).

Számos, több nyelvet támogató modell is megjelent, amelyeknek támogatott nyelvei között a magyar is szerepel. Ilyen modellek például az mBERT (Devlin és mtsai, 2018), XLM-RoBERTa (Conneau és mtsai, 2020) és RemBERT (Chung és mtsai, 2020). Az mBERT egy klasszikus BERT modell, ami 104 nyelven lett tanítva. Az XLM-RoBERTa, a RoBERTa (Liu és mtsai, 2019) előtanítási mechanizmusát követő, 100 nyelvre betanított modell. A RemBERT esetén a be- és kimeneti beágyazások szétválasztására tettek javaslatot (általában ezek a súlyok meg vannak osztva egymás között), és külön dimenzióméreteket rendelnek azokhoz, ezzel a generalizációját a modellnek tudják befolyásolni. Ez a modell rendelkezik azzal a jó tulajdonsággal, hogy a finomhangolás során a paraméterek száma jelentősen csökkenthető a kimeneti beágyazás eldobásával.

Az évek során számos arany és ezüst minőségű annotációval ellátott adathalmaz vált elérhetővé magyar nyelvre. Névelemfelismeréshez a SzegedNER (Szarvas és mtsai, 2006), NerKor (Simon és Vadász, 2021) és a NerKor 1.41e (Novák és Novák, 2022), szekvenciaosztályozáshoz egyebek mellett az OpinHuBank (Miháltz, 2013) egy gyakran használt adathalmaz, amelyben egyes személyekhez kapcsolódó szentimentkategóriát kell prediktálni, azok szövegkörnyezetének

[1] `https://huggingface.co/SzegedAI`

függvényében. További kiértékelési lehetőségeket biztosít a HuLU (Ligeti-Nagy és mtsai, 2022) (Hungarian Language Understanding Benchmark Kit), ami 6 adathalmazt tartalmaz.

A magyar nyelvi erőforrások kiértékelésére több munka is született már. Ács és mtsai (2021) a HuBERT modellt vetette össze 4 további többnyelvű modell használatával a névelemfelismerési, szófaji és morfológiai egyértelműsítési feladatokon. Az egyes transzformerek teljesítményét rétegekre bontva is megvizsgálták, ami elég nagy mélységet ad az eredményeknek. Simon és mtsai (2022) a NerKor és SzegedNER korpuszok alkalmazását járja körbe különböző ipari minőségű nyelvi keretrendszer (vagy modell) alkalmazása mellett.

3. HuBERTUSz előtanítása

A BERT-jellegű transzformer modellek hiperparamétereinek megválasztásával, valamint az előtanítás pontos menetével kapcsolatosan változatos – és időnként hiányosan dokumentált – megoldásokkal lehet találkozni az irodalomban. Munkánkban, a transzformer modelljeink létrehozása során elsődlegesen a Nemeskey (2020) által alkalmazott módszertant követtük.

Erőforráskorlátok miatt tanítókorpuszként a Hungarian WebCorpus 2.0 Wikipedia alkorpuszára támaszkodtunk. A számítási igény mérséklésére azt az általánosan alkalmazott stratégiát követtük, amely az előtanítást két fázisra osztja úgy, hogy a lépések első 90%-ában redukált, legfeljebb 128 szubtokennyi hosszal rendelkező inputok feldolgozása történik meg, míg a tanítás második, befejező fázisában az inputok hossza elérheti az 512 szubtokennyi hosszt is akár. A szöveg tokenizálására a HuBERT szótárát[2] használjuk fel.

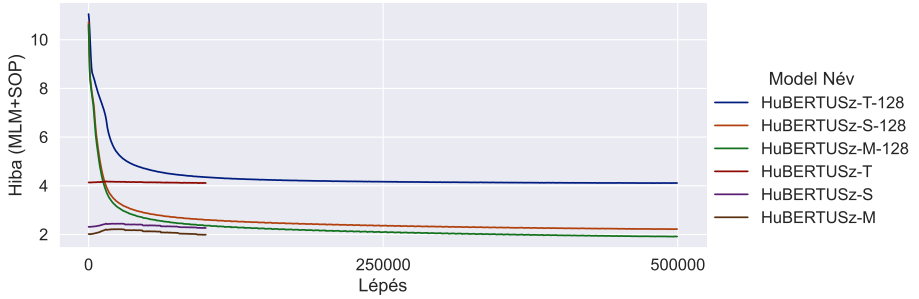
Liu és mtsai (2019) rámutattak, hogy a dinamikusan maszkolt nyelvimodellezés (MLM) – a BERT tanításánál is alkalmazott statikus párjával szemben – jobb általánosítóképességgel rendelkezik. Ennél fogva mi is ezt követjük, ami az alacsonyabb paraméterszámú modelljeink teljesítményét javíthatja, kisebb adathalmazon történő előtanításhoz mérve.

Lan és mtsai (2019) is rámutatott arra, hogy az eddig megszokott Next Sentence Prediction (NSP) feladat túl egyszerű a modell számára. Mivel az NSP során a feladat egy egyszerű témafelismerési feladatként viselkedik, helyette a Sentence Order Prediction-re (SOP) tesznek javaslatot, ami jobban képes a mondatok közötti koherenciára fókuszálni.

Az általunk tanított modellek dinamikus MLM-et és SOP-t oldanak meg, ahol a maszkolt nyelvi modellezés során alkalmazott maszkolási valószínűség megegyezik a BERT esetében alkalmazott értékkel ($p = 0,15$). A különböző modelljeink konfigurációi áttekinthetőek az 1. táblázat első 3 sorában. Három különböző méretű modellt (`tiny (T)`, `small (S)` és `medium (M)`) állítunk elő, amelyek checkpointjait megosztjuk az első és második fázis végéről is.

Az előtanítások során 64-es gradiens akkumuláció alkalmazásával 1024-es batchméretet értünk el, 10^{-4} tanulási rátával, 15000-es warmuppal, 500000 lépésen keresztül. A második fázisban a batch méretet 384-re állítottuk (batch =

[2] `https://huggingface.co/SZTAKI-HLT/hubert-base-cc`



1. ábra. A modellek előtanítása során a veszteségfüggvény alakulása. A `*-128` elnevezésű modellek az első fázis eredményeit, az ezen jelölést mellőző modellek pedig a második fázis modelleit jelölik.

16, gradiens akkumuláció = 24), és 100000 lépést hajtunk végre 512-es maximális szekvenciahossz mellett. You és mtsai (2019) ajánlása szerint a második fázisban ún. *re-warm-up*-ot alkalmazunk, vagyis a második fázis kezdetekor – az addigra alacsony értéket fölvevő – tanulási ráta értékét megemeljük. A tanításhoz használt kód elérhető GitHub-on[3], illetve a tanítás menete végigkövethető az 1. ábrán vagy ennél részletesebben a Weights&Biases[4] felületén.

4. Kísérletek

Az előállított modelljeinket számos feladaton finomhangoljuk. Az összes kísérlet részletes eredménye és konfigurációja megtekinthető a Weights&Biases[5] felületén, de a fontosabb paramétereket itt is megemlítjük. Továbbá mind az első (`HuBERTUSz-{T,S,M}-128`), mind a második (`HuBERTUSz-{T,S,M}`) előtanítási fázis által elkészült modelljeinkün is elvégezzük kiértékeléseinket.

4.1. Modellek

A lehetséges modellek kiválasztása során törekedtünk a lehető legjobban lefedni a könnyen hozzáférhető magyar nyelvet támogató modellek minél szélesebb spektrumát. Ami így tartalmaz már használt és újonnan megjelentetett modelleket is, amelyekről magyar nyelv kapcsán nem történtek eddig mérések. A választott modellek listája és a hozzájuk tartozó paraméterek áttekinthetőek az 1. táblázatban.

Egynyelvű modellek A már baseline-nak tekinthető HuBERT-et (Nemeskey, 2020, 2021) vesszük kísérleteink alapjául. Ehhez a *hubert-base-cc* modellt

[3] `https://github.com/ficstamas/hubertusz-pretraining`
[4] `https://wandb.ai/szegedai-semantics/hubertusz-pretraining`
[5] `https://wandb.ai/szegedai-semantics/hubertusz-finetuning`

Név	#Nyelvek	$\|\theta\|$	$\|\theta_{emb}\|$	Méret	$\|V\|$	#Rétegek	#Fejek	h_{hidden}	h_{inter}
HuBERTUSz-tiny	1	4,5 M	4,1 M	0,04x	32001	2	2	128	512
HuBERTUSz-small	1	29,5 M	16,6 M	0,26x	32001	4	8	512	2048
HuBERTUSz-medium	1	42,1 M	16,6 M	0,38x	32001	8	8	512	2048
HuBERT-base-wiki	1	109,4 M	23,8 M	0,99x	30501	12	12	768	3072
HuBERT	1	110,6 M	24,9 M	1,00x	32001	12	12	768	3072
XLM-RoBERTa-base	100	278,0 M	192,3 M	2,51x	250002	12	12	768	3072
XLM-RoBERTa-large	100	559,8 M	256,5 M	5,06x	250002	24	16	1024	4096
mBERT	104	177,8 M	92,2 M	1,60x	119547	12	12	768	3072
distil-mBERT	104	135,3 M	92,2 M	1,22x	119547	6	12	768	3072
RemBERT	110	575,9 M	64,2 M	5,20x	250300	32	18	1152	4608

1. táblázat. A használt modellek paramétereinek összevetése. θ a paramétereket, V a szótárat, h_{hidden} és h_{inter} a rejtett és köztes reprezentációk dimenzióját jelölik. A Méret oszlopban a modellek HuBERT-hez viszonyított relatív mérete található.

használjuk (`HuBERT`), amely HuggingFace-ről könnyedén elérhető, és a teljes Magyar Webcorpus 2.0-on lett tanítva. Ezen túl a ritkábban használt, kizárólag a Webcorpus 2.0 Wikipedia alkorpuszán tanított `HuBERT-wiki-cased` és `HuBERT-wiki-uncased` variánsokat is bevesszük a kísérleteinkbe a teljesség kedvéért, illetve mivel az általunk létrehozott modellek is ugyanezen szövegeken lettek előtanítva.

Többnyelvű modellek Az `mBERT` a legelső többnyelvű transzformerek egyike, amely a hagyományos BERT tanítási stratégiával lett előállítva 104 nyelven. A `distil-mBERT` az `mBERT` egy disztillált (Sanh és mtsai, 2019) variánsa, ahol a rétegek számát lefelezték. Az `XLM-RoBERTa` a RoBERTa (Liu és mtsai, 2019) által bemutatott tanítási stratégiával előállított többnyelvű transzformer, amelyet Conneau és mtsai (2020) mutatott be. A `RemBERT` (Chung és mtsai, 2020) létrehozása során szétválasztották a be- és kimeneti beágyazásokat. Ezek dimenzióinak megfelelő allokálásával a paraméterek számosságát tudták csökkenteni a finomhangolási lépéshez. Ez úgy valósulhatott meg, hogy kisebb dimenziószámot adtak a bemeneti beágyazási rétegnek, míg az MLM fejben található kimeneti beágyazásnak ennél több paraméter jutott. Mindennek köszönhetően több réteg és fej használatára nyílik lehetőség ennél a modellnél.

4.2. Adathalmazok

Modelljeink kiértékelésére diverz feladatokat választottunk, jól definiált adathalmazokon, amelyeken a modellek viselkedése jól összehasonlítható. Ezek a feladatok a névelemfelismerés, szófaji egyértelműsítés, szentimentosztályozás, valamint a természetesnyelvi következtetés különféle problémái.

Tokenklasszifikációs faladatok közül névelemfelismerésben és szófaji egyértelműsítésben vizsgáltuk meg a modelljeink viselkedését. A névelemfelismerés

terén két adathalmaz használata mellett döntöttünk, amelyeknek szöveges tartalma majnem teljesen átfed. Az egyik a NerKor (Simon és Vadász, 2021), ami hagyományos névelemfelismerési feladatot definiál, és a megszokott *PER*, *LOC*, *ORG* és *MISC* címkéket alkalmazza a fikciós (fiction), jogi (legal), hír (news), web és Wikipedia alkategóriák mentén. Novák és Novák (2022) a NerKor adathalmaz címkézésen javasolt módosításokat, ami eredményeképp a teljes adathalmazban így már 34 egyedi címke szerepel, amelyek a *OntoNotes 5.0* standardot követik. Ezen felül még egy autókról szóló részhalmazzal is bővítette a meglévő korpuszt. A szófaji egyértelműsítésre vonatkozó kiértékeléseinket a NerKor univerzális morfológiai információkkal ellátott alkorpuszain (hír, web és Wikipedia) hajtottuk végre.

Az OpinHuBank adathalmaz (Miháltz, 2013) személyek különböző kontextusban való előfordulásához rendel pozitív, negatív vagy neutrális címkéket. Az adathalmazt öten annotálták, ami alapján mi többségi döntés szerint hoztuk létre a végső címkét. Ha a többségi címke nem határozható meg egyértelműen holtverseny miatt, akkor az az adatpont neutrális címkét kapott.

A HuLU kiértékelő környezet (Ligeti-Nagy és mtsai, 2022) a már angolból is megszokott NLI (Natural Language Inference) feladatokat ülteti át magyar nyelvre, és az alábbi 6 adathalmaz alkotja:

i) a COLA, ami a mondatok nyelvhelyesség szerinti kategorizálását tűzi ki célul,
ii) a CoPA, ami egy feleletválasztós teszt,
iii) az SST2, ami egy szentimentosztályozási feladat,
iv) a WNLI, ami egy anaforafeloldási probléma,
v) a WS, ami egy anaforafeloldás, kérdésmegválaszolási alapon,
vi) az RC, ami pedig egy szövegértési problémát fogalmaz meg.

5. Eredmények

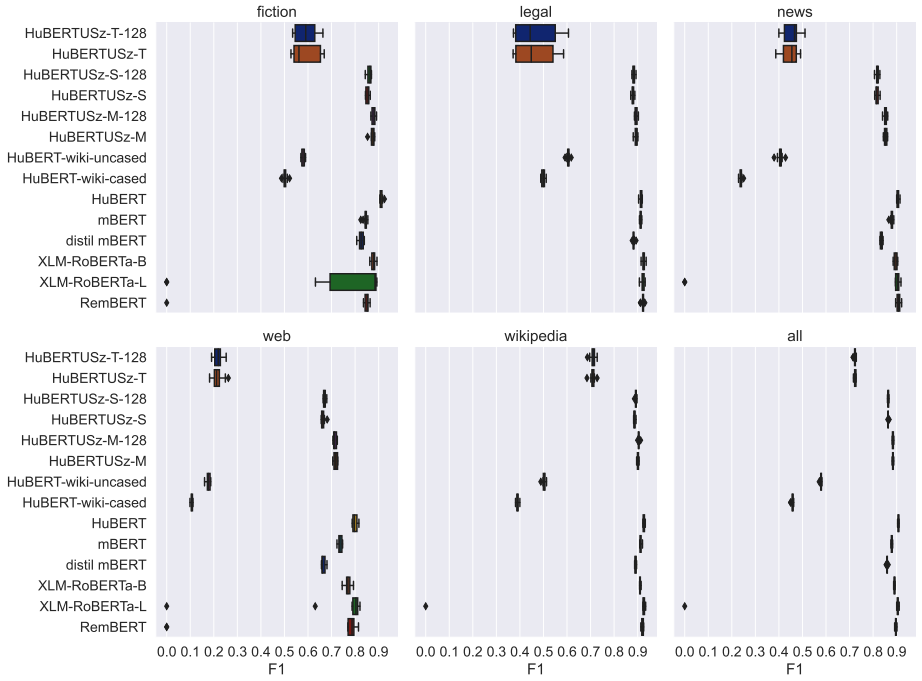
Az összes finomhangolásos kísérletet az osztályozó fejek 10 különböző random inicializálása mellett hajtottuk végre. Minden kísérlethez a következő hiperparamétereket alkalmaztuk: $5 \cdot 10^{-5}$ tanulási ráta, 32-es batch méret és 3 epoch. Minden más paraméter a `transformers` csomagban található *Trainer* alapvető paraméterének tekinthető, amelyek a már említett Weights&Biases oldalon is megtalálhatók.

A finomhangoláshoz és kiértékeléshez használt kódot a GitHubon[6] tettük elérhetővé, ami a használt adathalmazokat és modelleket HuggingFace kompatibilissá konvertálja futás közben. Ezenfelül a tokenszintű osztályozáshoz a metrikákat a *seqeval* könyvtár biztosította.

5.1. Finomhangolás – Névelemfelismerés

NerKor A hagyományos névelemfelismerési feladat eredményei alkorpuszonkénti lebontásban megtekinthetők a 2. ábrán. Az összes modell relatív teljesítménye

[6] `https://github.com/ficstamas/hu-eval/tree/mszny2023`



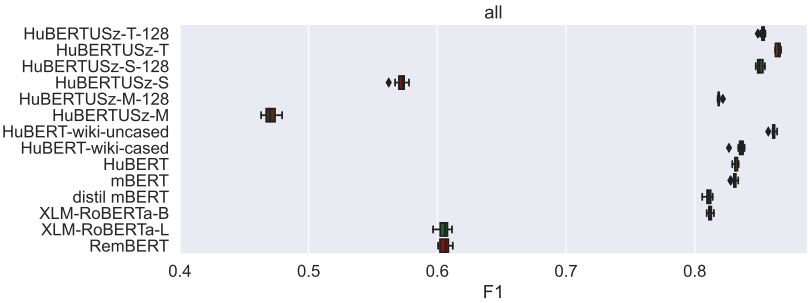
2. ábra. Finomhangolt modellek teljesítménye a NerKor egyes részhalmazain. Az *all* az egyes részhalmazok konkatenációját jelöli.

a különböző részhalmazokon lényegében azonos. Egyedül az `XLMRoBERTa-large` rendelkezik kiugró, outlier értékekkel. Az összes többi modell teljesítménye a megismételt kísérletek mentén stabilnak tekinthető. A legjobban eredményekkel rendelkező modellek között az F1-értékek különbsége szinte elhanyagolható, főleg ha figyelembe vesszük, hogy az egyes modellek hány paraméterrel rendelkeznek. Alacsony paraméterszámuk ellenére a `HuBERTUSz` modellek relatív teljesítménye jónak mondható, hiszen a `HuBERT`-ben lévő paramétereknek csupán a töredékével rendelkeznek (a `HuBERT` mérete kb. 2,6-szorosa a *medium* modellünknek).

NerKor 1.41e: Az adathalmazhoz tartozó összes alkorpusz aggregálása mentén kapott eredmények a 3. ábrán láthatók.

Megfigyelhetjük, hogy ezen feladat esetén nem feltétlen a paraméterek mennyisége számít, hiszen a *small* és *medium* konfigurációink a paraméterek mennyiségét tekintve relatív jó teljesítményt érnek el a többi modellhez képest. A végső sorrend a magyar nyelvű modellek esetén azonban a paraméterek mennyiségét követi.

A multilinguális modellek esetén hasonló viszonyokat nehéz leszűrni, mert mindegyik a paramétereit különböző módon osztja el. A végső sorrend sem ezek

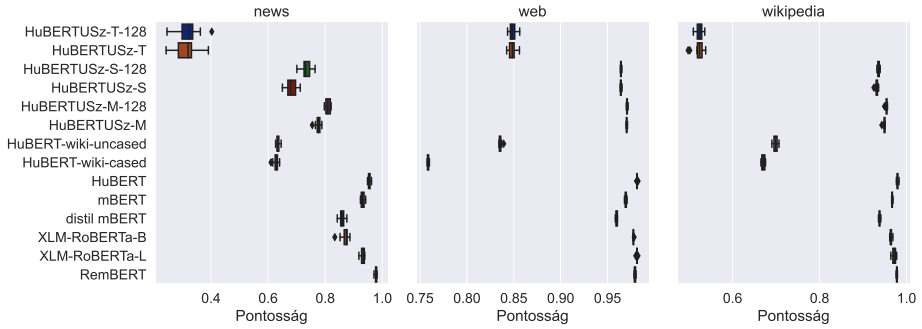


3. ábra. Finomhangolt modellek teljesítménye a NerKor 1.41e adathalmazon, ahol az *all* halmazt az egyes részhalmazok konkatenációjával kaptuk. A részhalmazonkénti eredmények megtekinthetők a Függelék 8. ábráján.

mennyiségével arányos. Azt viszont kijelenthetjük, hogy a `XLM-RoBERTa-large` modell teljesít a legjobban az összes modell közül.

5.2. Finomhangolás – Szófaji egyértelműsítés

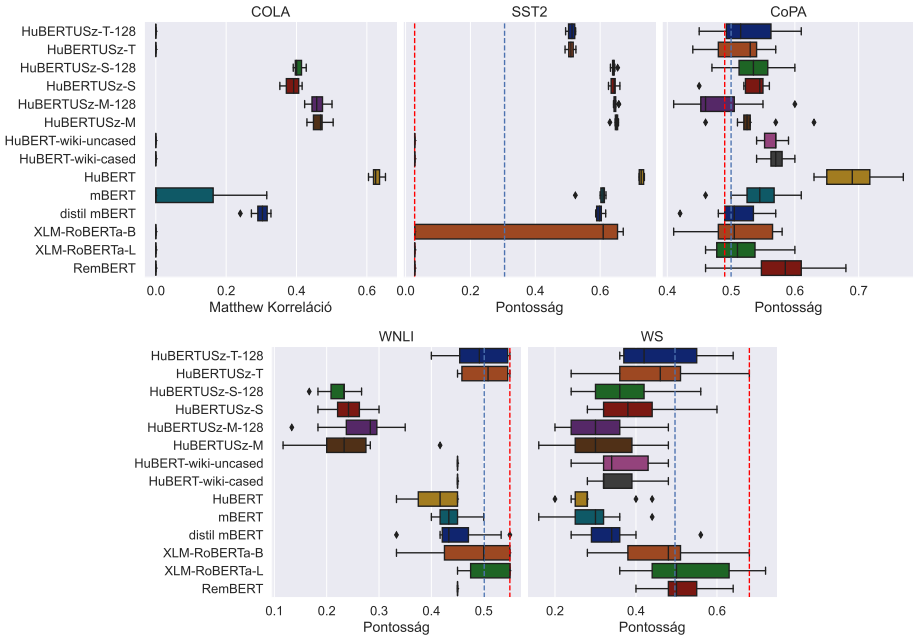
A 4. ábrán látható, hogy a legtöbb modell magas pontossággal oldotta meg a szófaji egyértelműsítési feladatot. A különböző modellek eredménye között a *news* adathalmazon figyelhető meg a legnagyobb szórás, ami valószínűleg az alkorpusz méretéből (is) adódik, mivel mindösszesen 9178 címkézett szóból áll (a vágások között összesen). Ezzel szemben a *web* és a *wikipedia* részhalmazok rendre 188250, illetve 26764 címkézett szóelemmel rendelkeznek.



4. ábra. A finomhangolt modellek teljesítménye a NerKor egyes részhalmazain a szófaji egyértelműsítési feladaton.

5.3. Finomhangolás – HuLU

A különböző modellek HuLU feladatokon nyújtott eredményességét az 5. ábrán láthatjuk. Piros szaggatott vonallal jelöltük annak a baseline modellnek a teljesítményét, ami minden esetben azonosan a tanítóhalmazon többségben lévő címkét prediktálja. A kék szaggatott vonal a véletlenszerű döntést hozó modell teljesítményét jelöli. Ezek a *COLA* esetén 0-nak tekinthetőek a Matthew korrelácóis együttható viselkedése miatt.



5. ábra. A HuLU kiértékelésre használt feladatain kapott eredmények. A szaggatott piros vonal a többségi, a szaggatott kék vonal pedig a véletlen baseline eredményét jelöli. A COLA feladaton ezek értéke azonosan 0-t vesznek fel a Matthew korrelációs együttható viselkedése okán.

Az egyes feladatokra vonatkozó vágások és az azokon belüli címkék eloszlásai megtekinthetők a 2. táblázatban. Minden eredménynél érdemes figyelembe venni az adathalmazok méretét és a címkék eloszlását, amelyek alapján a *COLA* tekinthető a legreprezentatívabb feladatnak a vizsgáltak közül.

A *COLA* esetén kapott eredmények alapján egyértelműen kijelenthető, hogy a többnyelvű modellek nem képesek kezelni ezt a feladatot. Még úgy sem, hogy némelyiknek 5-ször annyi paraméter áll rendelkezésére, mint a `HuBERT`-nek. Ez a viselkedés potenciálisan javítható jobb hiperparaméterválasztással, azonban ez kifejezetten erőforrásigényes lenne ennyi modell mentén. Érdekes módon a

	COLA	CoPA	SST2	WNLI	WS
# Tanító adathalmaz	7274	400	9328	562	170
# Kiértékelő adathalmaz	910	100	1165	60	25
Címkeeloszlás a tanító adathalmazon	22:78	50:50	31:29:40	51:49	51:49
Címkeeloszlás a kiértékelő adathalmazon	22:78	49:51	60:37:3	55:45	32:68

2. táblázat. A HuLU általunk használt feladatainak statisztikái.

kizárólag a Wikipedián tanult `HuBERT` modellek nem viselkednek túl jól se ezen, se a többi feladaton. A `HuBERT` és `HuBERTUSz` modellvariánsok a paramétereik számával arányosan követik egymást.

Az *SST2* esetében a *COLA* feladatnál látottakhoz hasonló viselkedéseket figyelhetünk meg. Ennél a feladatnál azonban a címkék eloszlása meglehetősen érdekesen alakul: a többségi döntés alapján a modellek $\approx 3\%$-ot tudnak elérni csupán, és ez is figyelhető meg azokban az esetekben, amikor 0 közelinek látszanak a felvett értékek. Ismételten a `HuBERT` teljesít a legjobban, ugyanakkor – a legkisebb *tiny* modellt leszámítva – az alacsonyabb paraméterszámmal rendelkező `HuBERTUSz` modellek eredményei sem sokkal maradnak el.

A CoPA esetén a `HuBERT` teljesítménye kiemelkedő, míg az összes többi modell a bennük található paraméterszámtól függetlenül hasonló, a véletlenszerű tippelés eredményét nem sokkal meghaladó eredménnyel rendelkezik.

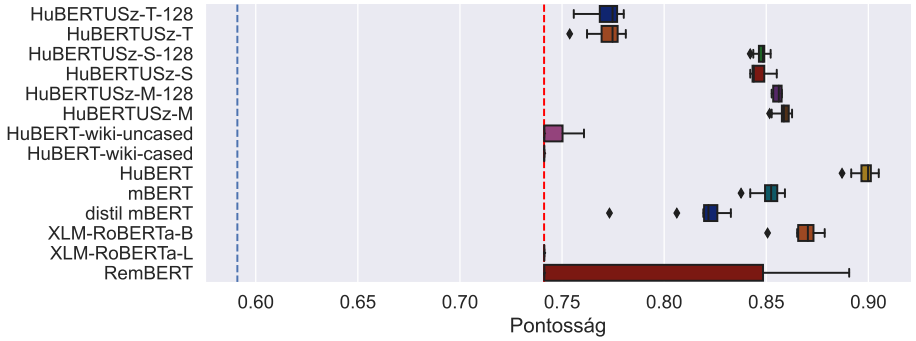
Az anaforafeloldáshoz kapcsolódó feladatokon (*WNLI* és *WS*) semelyik modell nem volt képest a baseline megoldások eredményét meghaladni. Mindez a modellek képességein és a feladatok nehézségén túl a rendelkezésre álló adatok méretével és minőségével is összefügghet, hiszen például a *WNLI* néhány tucatnyi tesztmondatai között angol nyelvűek is szerepeltek. Az előbbiek miatt, az anaforafeloldási feladatokon kapott eredmények inkább a teljesség igényét szolgálják, azokból messzemenő konklúziót nem vonnánk le.

Összességében a feladatok megoldására az egynyelvű modellek a vártak szerint viselkedtek (leszámítva a kizárólag a Wikipedián előtanított `HuBERT` modellek gyenge teljesítményét a *COLA* és *SST2* feladatokon), azonban meglepő módon a többnyelvű társaik nem voltak képesek használható szemantikai jellemzőket kinyerni.

5.4. Finomhangolás – Szentimentosztályozás

Az OpinHuBankon elért eredményeket a 6. ábra tartalmazza, amelyről leolvasható, hogy sem az `XLM-RoBERTa-large`, sem pedig az `HuBERT-wiki-*` modellek nem voltak képesek a többségi döntést hozó baseline eredményét érdemben meghaladni, így ezen feladaton az ezekre a modellekre építő modellalkotási kísérleteinket sikertelennek tekinthetjük.

Általánosságban a szentimentosztályozáson az egynyelvű modellek eredménye a paraméterek számával együtt növekszik (leszámítva a kizárólagosan Wikipedián előtanított `HuBERT` modelleket). A többnyelvű modellek esetén a `RemBERT`



6. ábra. Az OpinHuBankon elért eredmények.A függőleges szaggatott piros vonal a többségi, a kék a véletlen döntéssel rendelkező baseline eredményét mutatja.

és az `mBERT` eredményessége komoly eltérést mutat, a `RemBERT` modell eredményeit emellett nagyfokú instabilitás is jellemzi.

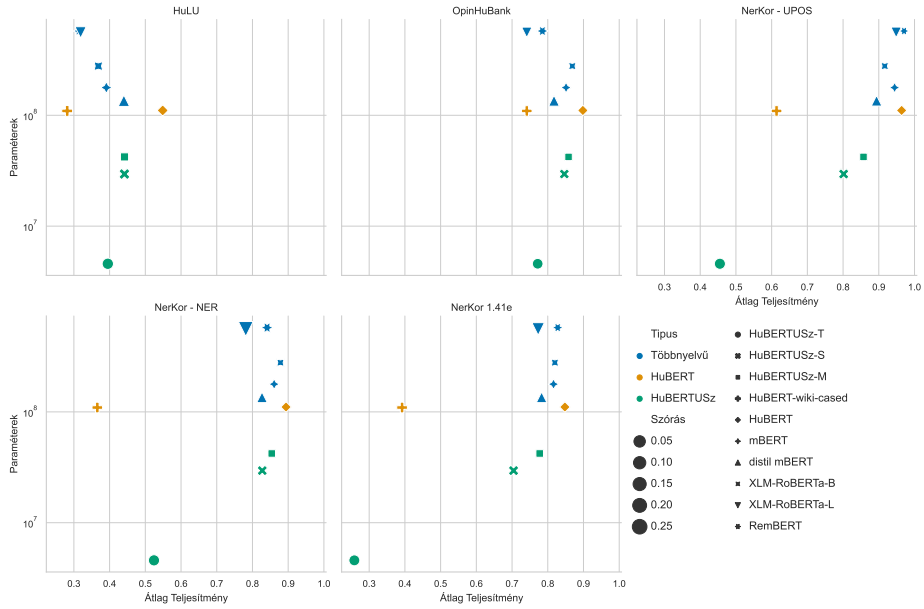
5.5. Paraméterek és teljesítmény viszonya

A 7. ábrán látható a paraméterek mennyiségének és az átlagos teljesítménynek a viszonya, ahol a szórást és átlagot az egyes részhalmazok között néztük. Az egyes részhalmazokon belül pedig a véletlen inicializációnkénti átlagot vettük.

Szekvenciaosztályozás esetén láthatjuk, hogy a *HuLU* megoldásához nem elég csak a paraméterek számának növelése, hanem ajánlott az egynyelvű modellekre támaszkodni. A *HuLU* esetén a `HuBERT` használata indokoltnak tűnik. Az *OpinHuBank* tekintetében is hasonló tendenciák figyelhetők meg, azonban ezen a feladaton már a `HuBERTUSz` modellek használata is érdemi alternatívaként tekintendő, hiszen a `HuBERT` úgy teljesít $\sim 4\%$-kal jobban a *medium* modellünkhöz képest, hogy közben több, mint 2,6-szer annyi paramétert használ. Megjegyzendő továbbá, hogy a – `HuBERTUSz` modellekhez hasonlóan – csak Wikipedia szövegeken előtanított `HuBERT` modellek teljesítményei jóval elmaradnak a `HuBERTUSz` modellek által kapottaktól.

A *NerKor* univerzális szófaj egyértelműsítési feladatán észrevehetjük, hogy a többnyelvű modellek is felzárkóztak. Ennek hátterében az állhat, hogy a feladat megoldásához általában elegendő a korai rétegek rejtett beágyazásai alapján döntést hoznunk. Más szóval élve a feladat a statikus szóreprezentációkhoz közel áll, amelyek a nyelv egyedi karakterkészlete miatt, a szótöredékek szinte csak a magyar nyelvet kódolják. Így nincsenek információval túlterhelve, mint a szekvenciaosztályozási token. A szófaji egyértelműsítési feladatoknál hasonló eredményeket fedezhetünk fel, amelynek hátterében hasonló okokat feltételezünk.

Összességében elmondható, hogy egy kisebb paraméterszámú egynyelvű modell jobb választásnak bizonyulhat egy többnyelvű modellhez képest. A `HuBERTUSz` modell pedig a `HuBERT`-hez képest elég közel kerül teljesítmény terén, figyelembe véve, hogy jóval kevesebb paraméterrel is rendelkezik.



7. ábra. A paraméterek és átlagos teljesítmény viszonya az egyes adathalmazokon. A szórást az egyes részhalmazok menti átlagos teljesítményből kaptuk.

6. Összegzés

Cikkünkben bemutattuk a legújabb előtanítással kapcsolatos ajánlásokat (dinamikus maszkolás, mondatsorrend-predikció) alkalmazó, eltérő méretekben előtanított `HuBERTUSz` modelljeinket, amelyeket a `HuBERT` esetében használtakhoz képest jóval kevesebb nyers adat és paraméter felhasználásával hoztunk létre. Az új modellek segítségével lehetőségünk nyílt a magyar nyelv vonatkozásában a különböző kapacitással rendelkező modellek skálázódásának összehasonlító vizsgálatára.

Kísérleteink a magyar nyelvet támogató, jóval nagyobb paraméterszámmal rendelkező többnyelvű modellek kiértékelésére is kiterjedtek. Összességében elmondható, hogy az általunk vizsgált feladatokon még a nagyságrendekkel kisebb kapacitással létrehozott egynyelvű modelljeink használata is kifizetődőbbnek bizonyult a többnyelvű modellek használatához képest, mind az elért eredmények, mind pedig az erőforrásigény szempontjából.

Az előállított modellek számos feladaton alkalmazhatónak tekinthetőek. Elsősorban szó szintű osztályozási feladatok megoldására ajánljuk, azonban itt is figyelni kell a megfelelő mennyiségű adatra.

Köszönetnyilvánítás

A kutatás az Emberi Erőforrások Minisztériuma ÚNKP-22-3 kódszámú Új Nemzeti Kiválóság Programjának támogatásával, valamint az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A MILAB SZTE nyelvtechnológia projekt nevében köszönetet mondunk az ELKH Cloud (lásd: Héder és mtsai (2022); `https://science-cloud.hu/`) használatáért, ami hozzájárult a publikált eredmények eléréséhez.

Hivatkozások

Ács, J., Lévai, D., Nemeskey, D.M., Kornai, A.: Evaluating contextualized language models for hungarian. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 15–28. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021)

Bhargava, P., Drozd, A., Rogers, A.: Generalization in NLI: ways (not) to go beyond simple heuristics. CoRR abs/2110.01518 (2021), `https://arxiv.org/abs/2110.01518`

Chung, H.W., Févry, T., Tsai, H., Johnson, M., Ruder, S.: Rethinking embedding coupling in pre-trained language models. CoRR abs/2010.12821 (2020), `https://arxiv.org/abs/2010.12821`

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020), `https://www.aclweb.org/anthology/2020.acl-main.747`

Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), `http://arxiv.org/abs/1810.04805`

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), `https://aclanthology.org/N19-1423`

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR abs/2010.11929 (2020), `https://arxiv.org/abs/2010.11929`

Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z.Gy., Váradi, T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 29–36. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021), `http://real.mtak.hu/120856/1/feldmann21.pdf`

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., Zhou, M.: Codebert: A pre-trained model for programming and natural languages. CoRR abs/2002.08155 (2020), https://arxiv.org/abs/2002.08155

Ficsor, T., Cserháti, R., Novák, A., Mihajlik, P., Zainkó, Cs., Berend, G.: Charmen electra - tokenizációmentes diszkriminatív nyelvi modellezés. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 45–58. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Héder, M., Rigó, E., Medgyesi, D., Lovas, R., Tenczer, S., Török, F., Farkas, A., Emődi, M., Kadlecsik, J., Mező, Gy., Pintér, Á., Kacsuk, P.: The past, present and future of the ELKH cloud. Információs Társadalom 22(2), 128 (aug 2022), https://doi.org/10.22503/inftars.xxii.2022.2.8

Hsu, W., Bolte, B., Tsai, Y.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. CoRR abs/2106.07447 (2021), https://arxiv.org/abs/2106.07447

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. CoRR abs/1909.11942 (2019), http://arxiv.org/abs/1909.11942

Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vadász, N., Yang, Z.Gy., Váradi, T.: HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 431–446. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019), http://arxiv.org/abs/1907.11692

McCoy, R.T., Pavlick, E., Linzen, T.: Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. CoRR abs/1902.01007 (2019), http://arxiv.org/abs/1902.01007

Miháltz, M.: OpinHuBank: szabadon hozzáférhetö annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 343–345. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2013), http://acta.bibl.u-szeged.hu/58859/1/msznykonf_009_343-345.pdf

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020), https://hlt.bme.hu/en/publ/nemeskey_2020

Nemeskey, D.M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021). pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021), http://acta.bibl.u-szeged.hu/73353/1/msznykonf_017_003-014.pdf

Novák, A., Novák, B.: NerKor 1.41e. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 389–412. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Orosz, Gy., Szántó, Zs., Berkecz, P., Szabó, G., Farkas, R.: HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 59–73. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108 (2019), http://arxiv.org/abs/1910.01108

Simon, E., Vadász, N.: Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In: Ekstein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12848, pp. 222–234. Springer (2021)

Simon, E., Vadász, N., Lévai, D., Nemeskey, D., Orosz, Gy., Szántó, Zs.: Az NYTK-NerKor több szempontú kiértékelése. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 403–416. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: Highly accurate named entity corpus for hungarian. In: International Conference on Language Resources and Evaluation. Genova (Italy) (2006)

Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962v2 (2019)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Yang, Z.Gy.: BARTerezzünk! messze, messze, messze a világtól, BART kísérleti modellek magyar nyelvre. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 15–29. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022a)

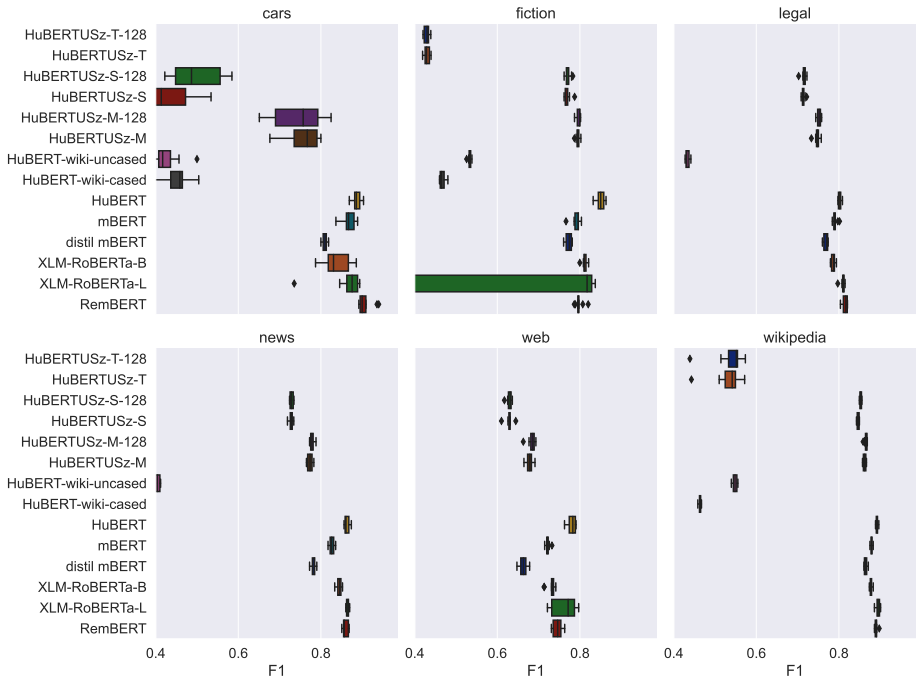
Yang, Z.Gy., Feldmann, Á., Váradi, T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő örnyezetben. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 603–617. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Yang, Z.Gy.: "Az invazív medvék nem tolerálják a suzukis agressziót" - Magyar GPT-2 kísérleti modell. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 463–476. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022b)

Yang, Z.Gy., Váradi, T.: Training language models with low resources: RoBERTa, BART and ELECTRA experimental models for hungarian. In: Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021). pp. 279–285. IEEE, Online (2021)

You, Y., Li, J., Hseu, J., Song, X., Demmel, J., Hsieh, C.: Reducing BERT pre-training time from 3 days to 76 minutes. CoRR abs/1904.00962 (2019), http://arxiv.org/abs/1904.00962

A. További eredmények



8. ábra. Finomhangolt modellek teljesítménye a NerKor 1.41e részhalmazain.