

HuBERTUSz: Alacsony paraméterszámú transzformer modellek létrehozása és kiértékelése magyar nyelvre

Ficsor Tamás Berend Gábor

Szegedi Tudományegyetem, Informatikai Intézet
[ficsort,berendg]@inf.u-szeged.hu

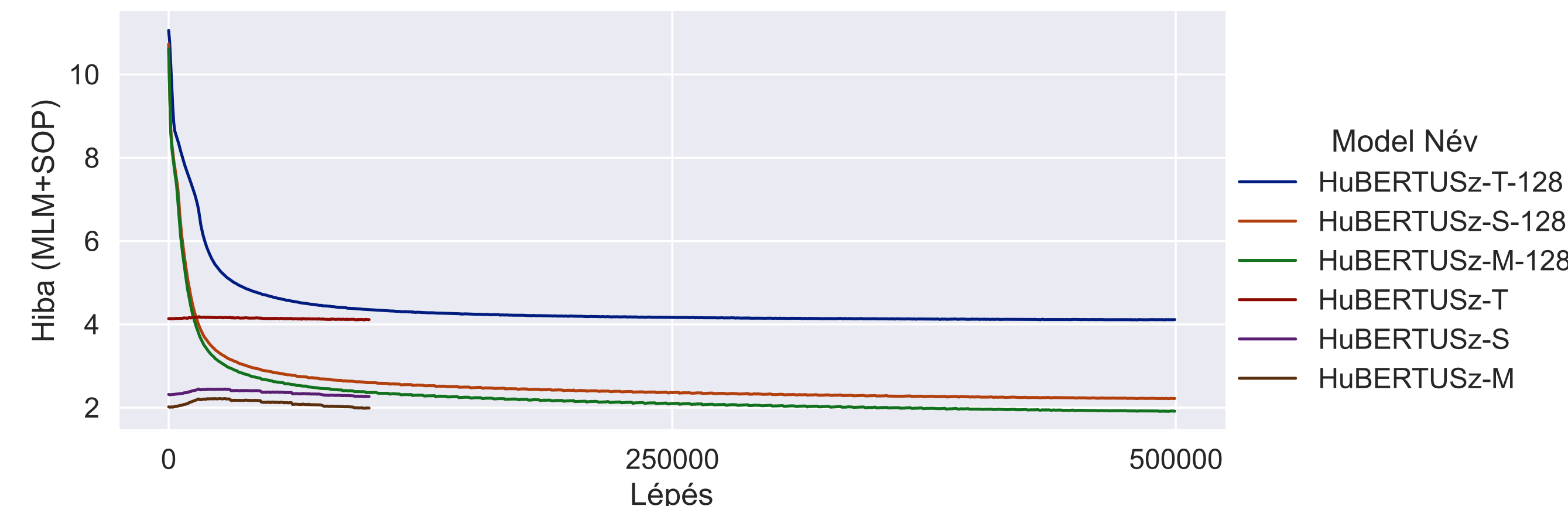
Bevezetés

Napjainkban a főbb hangsúly a nagy, több milliárd paraméterrel rendelkező modellekre esik. Azonban azok futtatása rengeteg erőforrást igényel. Sokkal kevesebb kísérlet irányul arra, hogy a kis paraméterrel rendelkező modellekkel a lehető legtöbbet ki tudjuk hozni. A már meglévő néhány magyar nyelvű és alacsony paraméterszámú modell kiegészítéseképp bemutatjuk a HuBERTUSz modelleket. A modellek létrehozása közben fontosnak tartottuk, hogy azokat a lehető legkönnyebben elérhetővé is tegyük, így megtalálhatók a HuggingFace felületén: <https://huggingface.co/SzegedAI>

Előtanítás

Az előtanítás folyamatát a meglévő kutatási eredményekhez igazítva, a lehető legeredményesebb módon igyekeztünk véghez vinni:

- A HuBERT (Nemeskey, 2020) tokenizáló modulját használtuk
- Adathalmaz: Hungarian WebCorpus 2.0, Wikipedia alkorpusza
- You és mtsai (2020) ajánlásának megfelelően 2 fázisú tanulást használtunk, rewarmup alkalmazása mellett
- Dinamikus maszkolást (Zhuang és mtsai, 2021) és mondatrend-meghatározást végeztünk (Lan és mtsai, 2020)



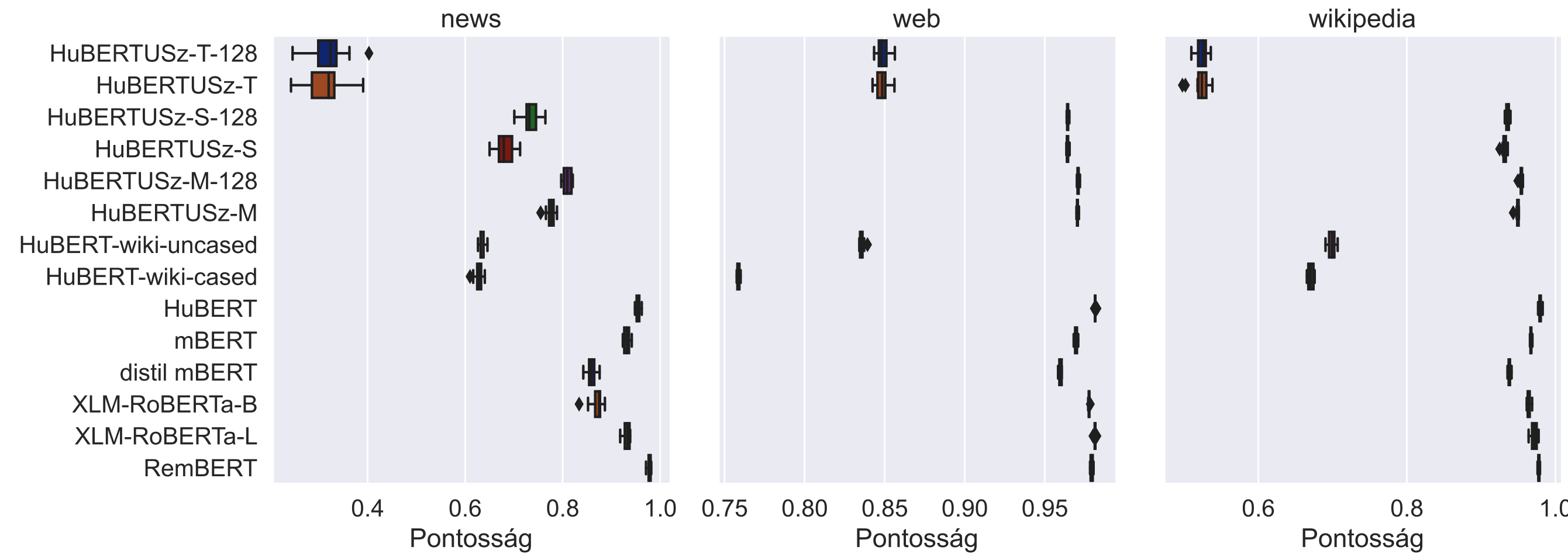
Név	#Nyelvek	$ \theta $	$ \theta_{emb} $	Méret	$ V $	#Rétegek	#Fejek	h_{hidden}	h_{inter}
HuBERTUSz-tiny	1	4,5 M	4,1 M	0,04x	32001	2	2	128	512
HuBERTUSz-small	1	29,5 M	16,6 M	0,26x	32001	4	8	512	2048
HuBERTUSz-medium	1	42,1 M	16,6 M	0,38x	32001	8	8	512	2048
HuBERT-base-wiki	1	109,4 M	23,8 M	0,99x	30501	12	12	768	3072
HuBERT	1	110,6 M	24,9 M	1,00x	32001	12	12	768	3072
XLNet-RoBERTa-base	100	278,0 M	192,3 M	2,51x	250002	12	12	768	3072
XLNet-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 szótárt, h_{hidden} és h_{inter} a rejtett és köztes reprezentációk dimenzióját jelölik. A Méret oszlopban a modellek HuBERT-hez viszonyított relatív mérete található.

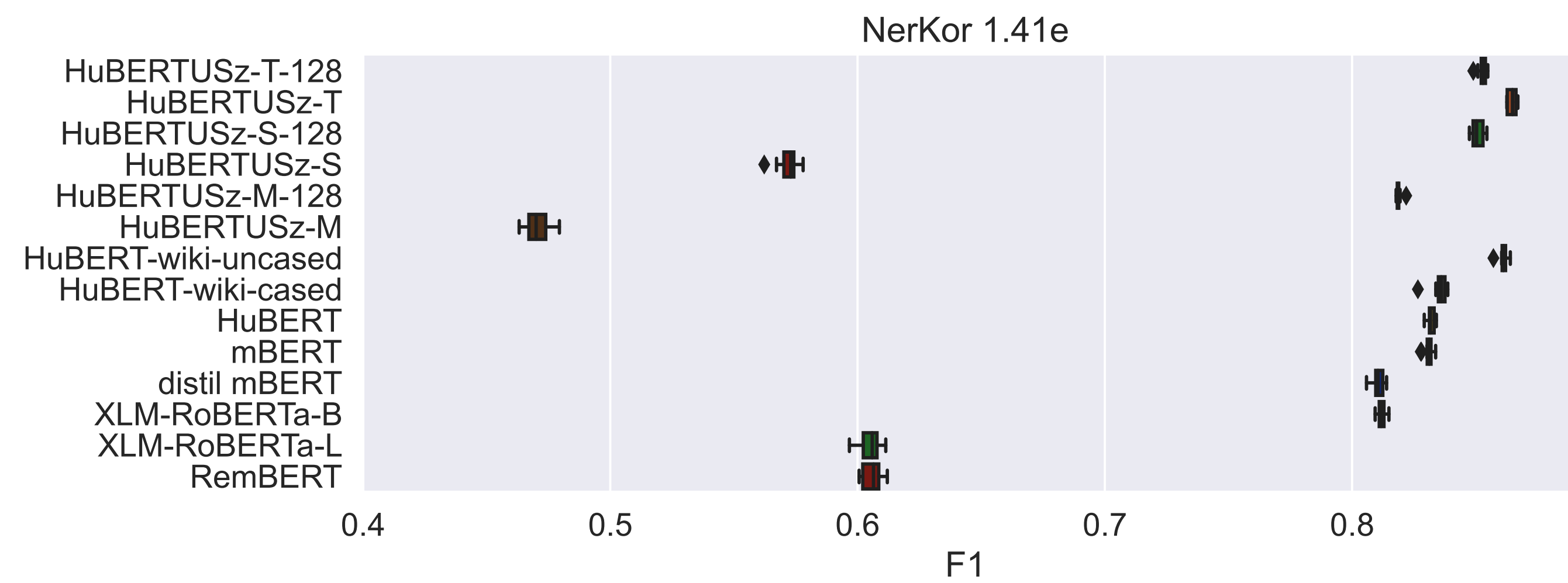
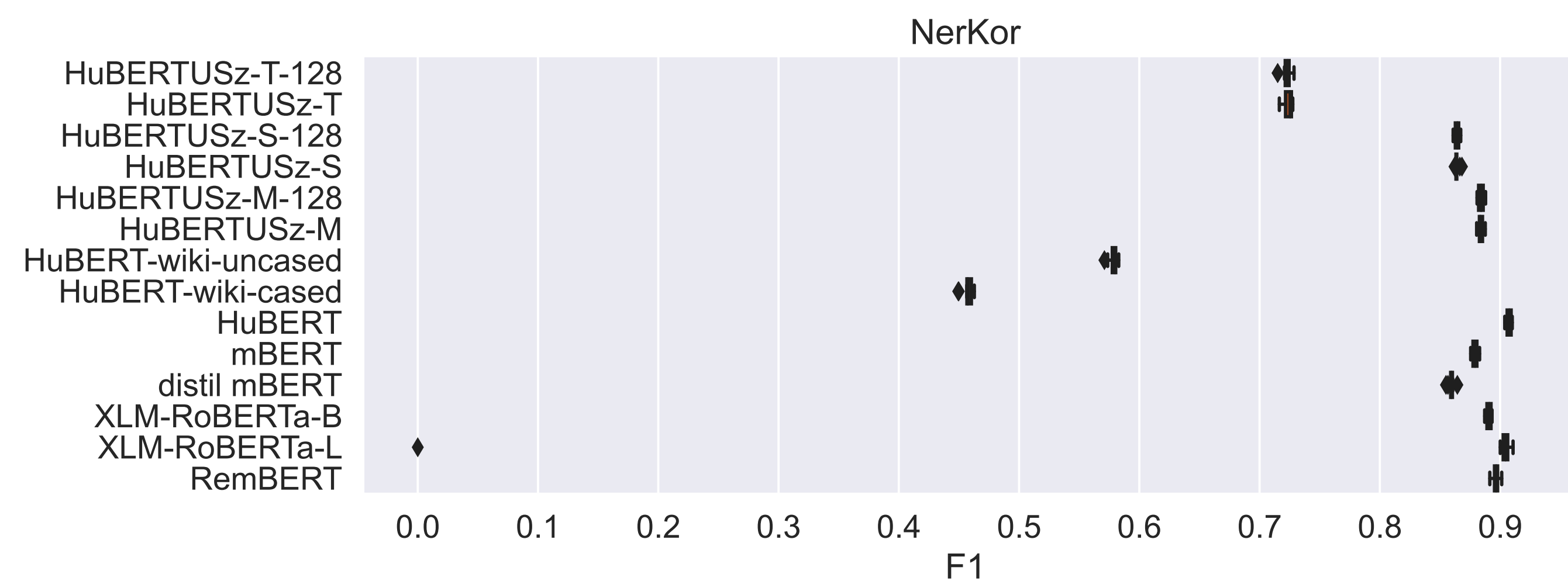
Kísérletek

- Szófaji egyértelműsítés: NerKor
- Névelemfelismerés: NerKor, NerKor 1.41e
- Szekvenciaosztályozás: HuLU (CoLA, SST2, CoPA, WNLI, WS), OpinHuBank

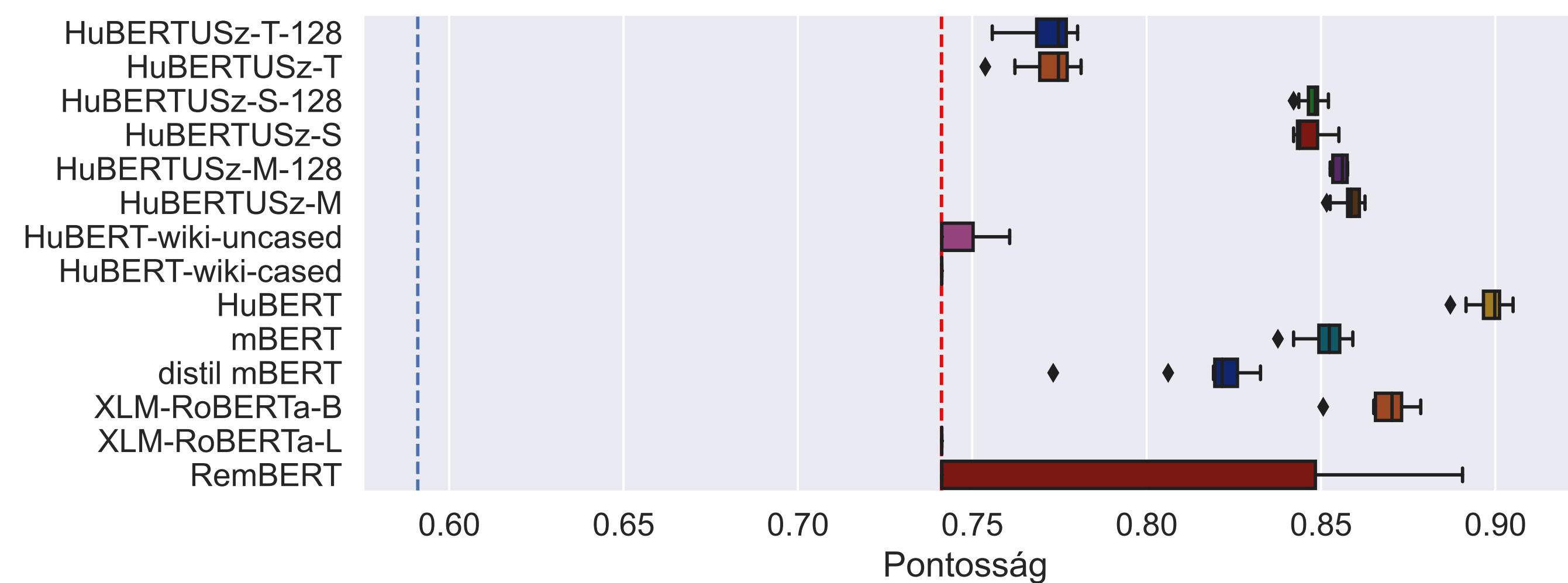
Szófaji egyértelműsítés



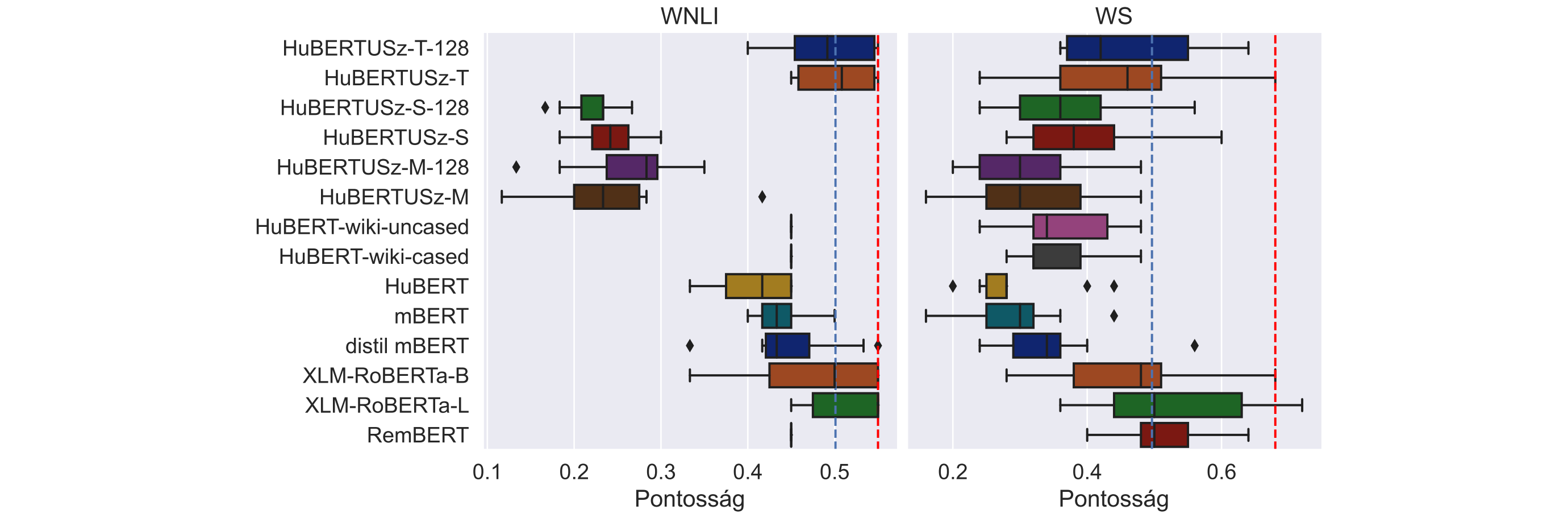
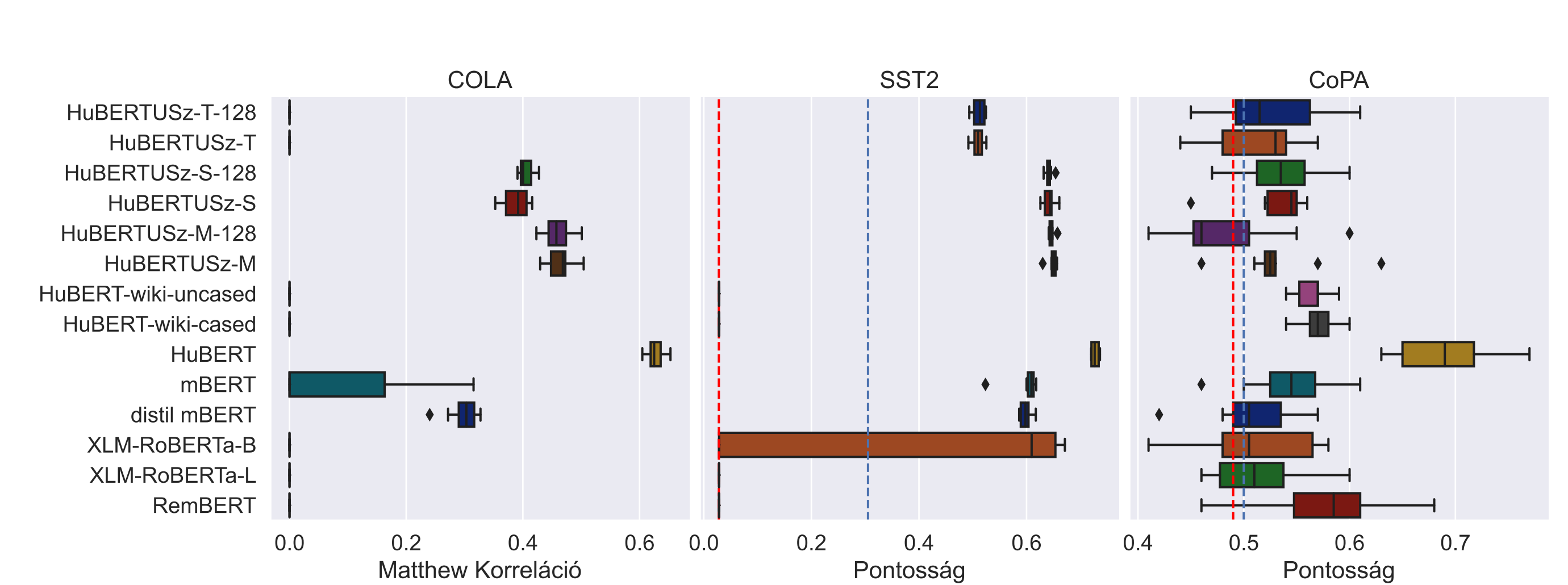
Névelemfelismerés



OpinHuBank



HuLU



Konklúzió

Általánosságban elmondható, hogy az elkészült modellek a paramétereik mennyiségével arányosan működnek. Az alacsony paraméterszámú modellek elsősorban tokenszintű osztályozás során működnek a legjobban. Szekvenciaosztályozás során pedig érzékenyek az adatok mennyiségére.

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 (<https://science-cloud.hu/>) használatáért, ami hozzájárult a publikált eredmények eléréséhez.

Hivatkozások

- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=H1eA7AEtvS>
- 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
- You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large batch optimization for deep learning: Training BERT in 76 minutes. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=Syx4wnEtvH>
- Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics, pp. 1218–1227. Chinese Information Processing Society of China, Huhhot, China (Aug 2021), <https://aclanthology.org/2021.cc1-1.108>