A SLICES társadalomtudományi alkalmazásai: Nagy nyelvi modellek finomhangolása a felhőben

Kovács Viktor, Bánóczy Martin

HUN-REN TK

Előzmények

- komparatív politikatudományi kutatások → szövegklasszifikáció
- példa: Comparative Agendas Project (CAP) kódkönyv
  - nemzetközi standard
  - 21 közpolitikai kategória (pl. *makrogazdaság, kultúrpolitika*)
  - 9 nyelv
  - 6 terület (pl. *média, jogalkotás*)
- kézi kódolás?
  - drága
  - időigényes
  - munkaszervezési, megbízhatósági problémák

Előzmények

- gépi kódolás?
  - szótáralapú megoldások (Albaugh és mtsai, 2013)
  - hagyományos gépi tanulás: Burscher, Vliegenthart, és De Vreese (2015); Karan és mtsai. (2016); Sebők, Kacsuk, és Máté (2022)
  - nagy nyelvi modellek (BERT): Frantzeskakis és Seeberg (2022)
  - ML → 0.6-0.7 F1
  - LLM → 0.7-0.9 F1
- hasonló feladatok:
  - Manifesto Project
  - szentiment, emócióelemzés

Megoldás: nagy nyelvi modellek

- SOTA: előtanított LLM + finomhangolás az adott feladatra
  - nincs szükség „hatalmas” tanítóadatra
  - többnyelvű modellek (pl. XLM-RoBERTa)
- cél: egy inferencia platform létrehozása kutatók számára, amely gyorsan és megbízhatóan képes végrehajtani a gépi kódolást többnyelvű szövegkorpuszokon nagy nyelvi modellek segítségével (*The Babel Machine*)
- a legegyszerűbb ötlet: teljes adaton hangolt többnyelvű modell
- de mi történik, ha több specifikus modellt tanítunk?
- hátráltató tényező: erőforrás (GPU)

Miért pályáztunk?

- A meglévő infrastruktúra nem volt elég...
- HUN-REN Cloud → 2 VM
  - V100 32GB
  - A100 40GB
- XLM-RoBERTa (nagyméretű modell)
- egynyelvű adaton hangolt modell → pár órától 1-2 napig
- teljes tanítóadaton hangolt modell → 4-5 nap!
- memória - batch méret limitációk
- nehezen párhuzamosítható folyamatok
- hiperparaméter hangolás?

Pályázás menete

- SLICES Open Call
- pályázati dokumentum:
  - rövid összefoglaló
  - melyik infrastruktúrához szeretnénk hozzáférni
  - a projekt részletes leírása és elvárt célok
  - módszertan, munkaterv
  - publikációk
  - tervek a jövőre nézve
  - physical vs remote access
- jelentkezés: február - eredmény: március
- havi riportok
- Experiment Feedback Report + SLICES-SC experiment repository

Eredmények (CAP)

- 9 nyelvspecifikus modell
- 41 nyelv-domain modell
- általános javulás

Table 2. Performance of Pooled vs. Language-Specific Models (weighted macro F1)

Language	Pooled Model	Language Specific Model
Danish	0.83	0.91
Dutch	0.79	0.83
Italian	0.71	0.75
Portuguese	0.71	0.89
French	0.63	0.71
German	0.61	0.69
English	0.59	0.84
Spanish	0.48	0.62
Hungarian	0.30	0.83

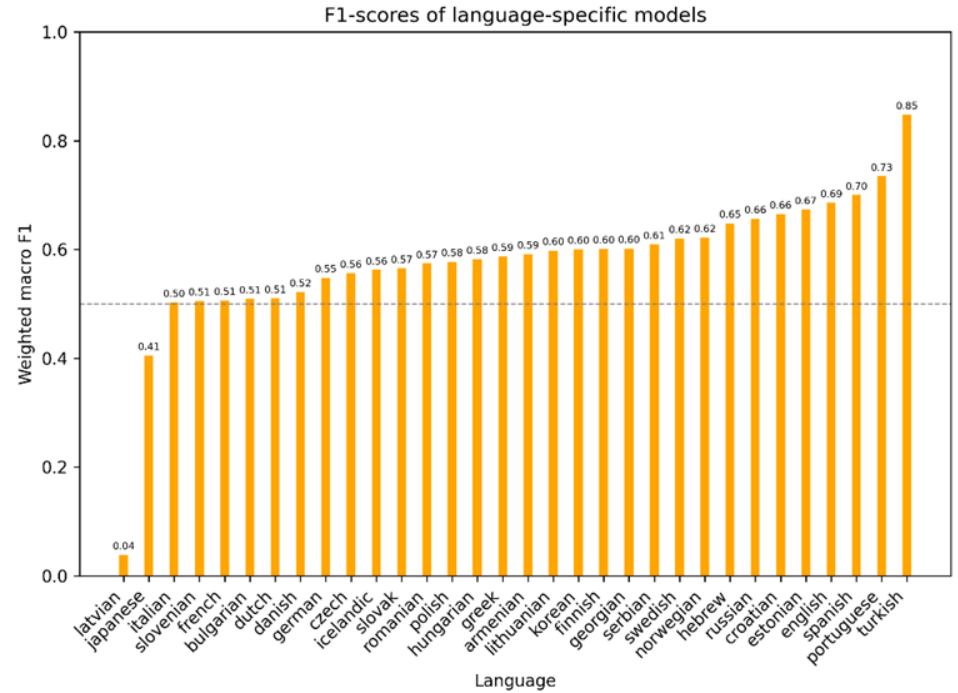
Table 3. Performance by Language-Domain Pairs (Weighted Macro F1)

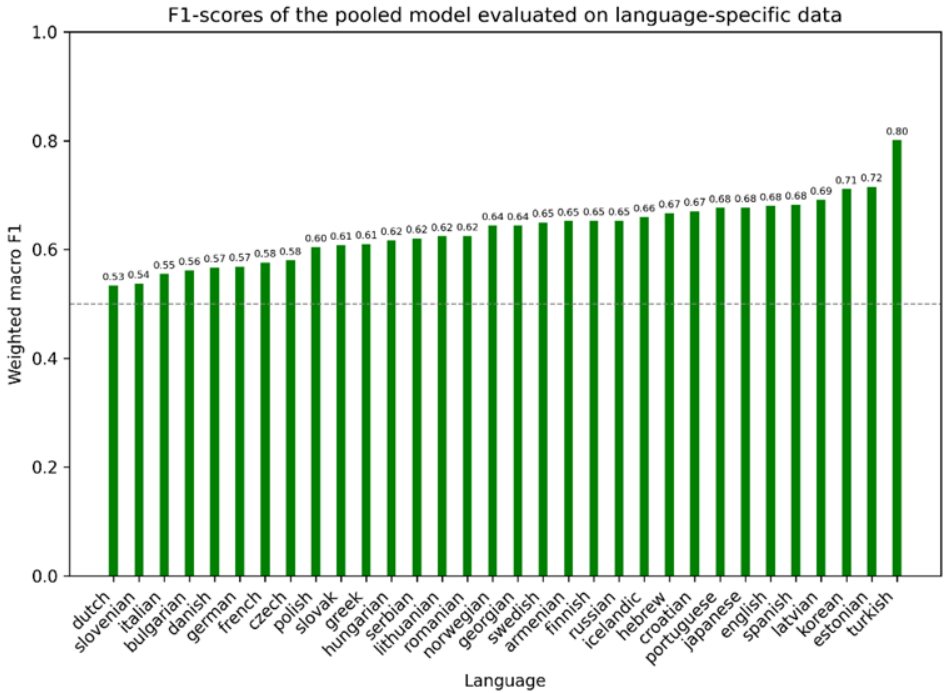
	Domain										
Language	Pooled Domain	Media	Social Media	Parl. Speech	Legisl.	Exec. Speech	Exec. Order	Party Manifesto	Judiciary	Budget	Public Opinion
Danish	0.91	-	-	0.94 (0.89)	0.86 (0.09)	0.63 (0.02)	-	-	-	-	-
Dutch	0.83	0.96 (0.46)	0.8 (0.16)	0.79 (0.09)	0.84 (0.16)	0.66 (0.06)	0.77 (0.07)	-	-	-	-
English	0.84	0.78 (0.16)	-	0.82 (0.01)	0.9 (0.65)	0.71 (0.07)	0.78 (0.04)	0.73 (0.06)	0.76 (0.01)	-	-
French	0.71	-	-	-	0.85 (0.21)	0.80 (0.17)	0.73 (0.09)	0.66 (0.53)	-	-	-
German	0.69	0.62 (0.08)	-	0.72 (0.1)	-	-	-	0.71 (0.83)	-	-	-
Hungarian	0.83	0.69 (0.07)	-	0.84 (0.73)	0.85 (0.01)	0.65 (0.1)	-	-	-	0.99 (0.08)	0.93 (0.00)
Italian	0.73	-	0.62 (0.32)	0.65 (0.29)	0.81 (0.39)	-	-	-	-	-	-
Portuguese	0.89	-	-	-	0.93 (0.53)	0.71 (0.09)	0.88 (0.38)	-	-	-	-
Spanish	0.62	0.76 (0.40)	-	0.38 (0.39)	0.85 (0.04)	0.71 (0.11)	0.85 (0.01)	0.75 (0.07)	-	-	-

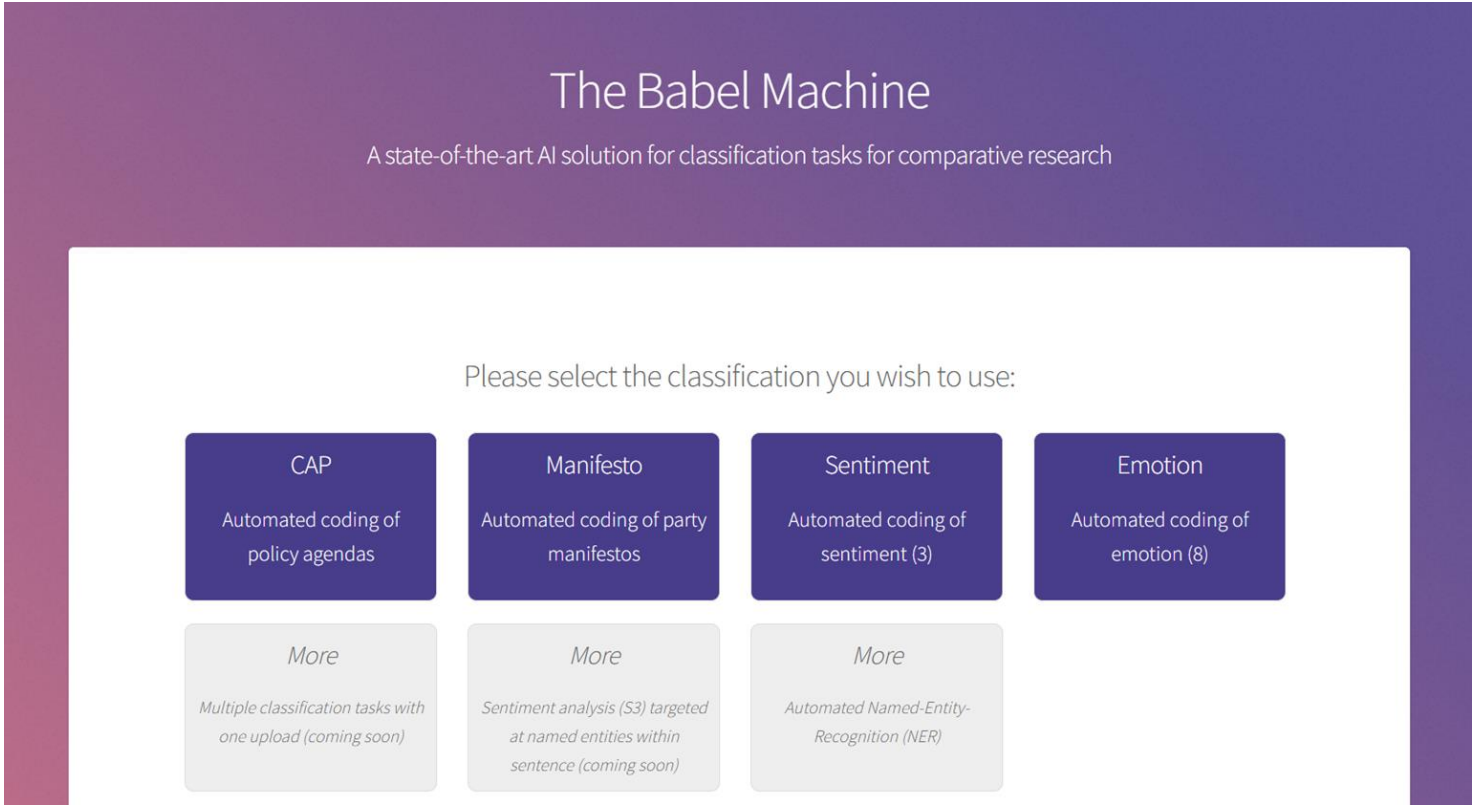
Note: The pooled-domain models in Column 2 are the same as the language-specific models in Table 2. The domains' share within each language corpora is in parenthesis. Small shares (e.g. 0.001) are rounded to 0.0.

Eredmények (Manifesto)

- 1 nagy többnyelvű modell - 32 „egynyelvű” modell







https://babel.poltextlab.com

Tapasztalatok

- a dokumentáció jó kiindulópont
- JupyterHub → development
- GPULab job → production
- elegendő tárhely (*project_antwerp*, *project_ghent*)
- sávszélesség optimális nagy adatok mozgatására

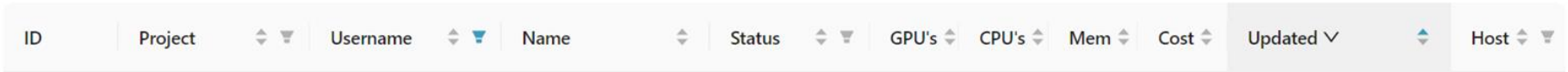
  (pl. modellek feltöltése HuggingFace-re)
- ad hoc feladatok: gépi fordítás, web scraping
- néha blocker: foglalt volt az adott GPU (de pár óra alatt sorra kerültünk)
- könnyű menedzselni, hogy kik férhetnek hozzá az adott projekthez

Hiányosságok?

- konténer tárhely betelik (nincs benne a dokumentációban)

The full message is:
The supervisor detected that this job's container used too much diskspace (10.46071836 GB).
This is not allowed, you need to configure storage like /project_ghent, /project_antwerp or /project_scratch to store data. The job has been stopped automatically.

- job list
  - vannak minimális szűrési és rendezési opciók
  - nem lehet rákeresni pl. konkrét dátumra
  - riportok elkészítésénél hasznos lett volna több funkció



- SLICES-SC Experiment Repository elérhetősége

Kezdeti nehézségek

- Workflow megértése
- Megfelelő klaszter kiválasztása (erőforrás rendelkezésre állás)
- jobDefinition json fájlok szintaxisának megértése
- Időbeli hatékonyság (ütemezések, értesítések)
- CPU - GPU kihasználása az adott feladat alapján

Köszönjük a figyelmet!

vkovacs@tk.hu
MartinBalazs.Banoczy@tk.hu