Hogyan segíti az MI az OPG termékek kategorizálását?

Kiskereskedelmi terméknevek kategorizálása termékcsoportokba mélytanulással

Putz Orsolya, PhD
BME, NAV MIMCS
putz.orsolya@tmit.bme.hu

NAV MIMCS

MIMCS = Mesterséges Intelligencia Munkacsoport

Elsődleges általános feladatok:

- Elkészíti a gépi tanuláshoz szükséges címkézési és validálási módszertant,
- gépi tanuló módszerek vizsgálata az adott területekre,
- gépi tanuló módszerek fejlesztése,
- értelmezi a gépi tanulás eredményeit,
- a NAV-val együttműködve prezentálja, publikálja, népszerűsíti az eredményeket.

A projektről

- Cél: Termékcsoportok árváltozásának idősoros elemzése, inflációs now-casting
- Kutatási irány:
  - Legújabb MI/mélytanulás alapú módszerek bevezetése a modellezésbe
  - GPU alapon
  - Fejlesztők: Putz Orsolya, Ónozó Lívia, Debreczeni Máté, Gyires-Tóth Bálint
  - Támogató szakértő: Járási István

- Jelenlegi megoldás:
  - Terméknevek kategorizálása TF-IDF módszerrel
  - CPU alapon
  - Vezető kutató, fejlesztő: Járási István

- Együttműködő szervezetek: NAV, NAV MIMCS, MNB, BME, SZTE, KSH

A projektről

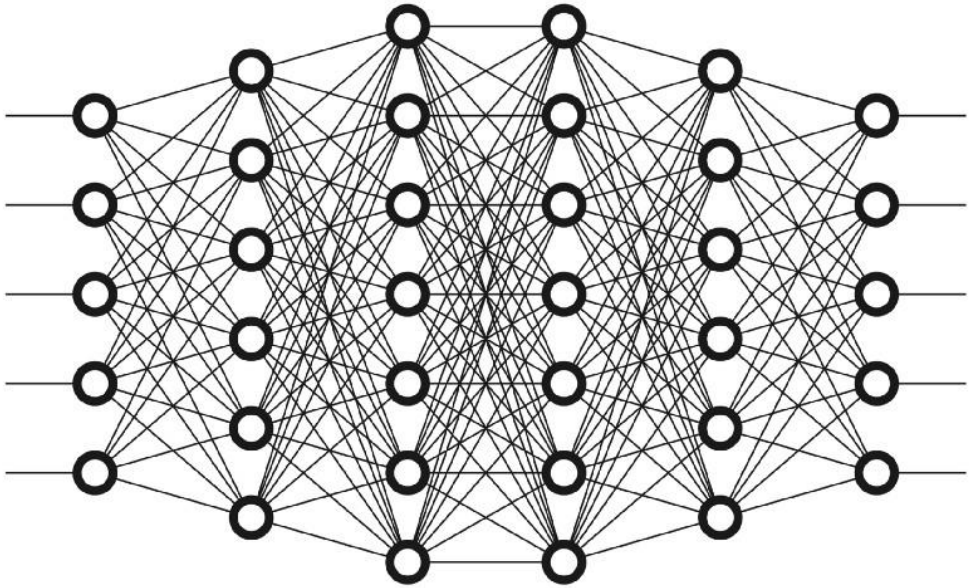
- Cél: Termékcsoportok árváltozásának idősoros elemzése, inflációs now-casting
- Kutatási irány:
  - Legújabb MI/mélytanulás alapú módszerek bevezetése a modellezésbe
  - GPU alapon
  - Fejlesztők: Putz Orsolya, Ónozó Lívia, Debreczeni Máté, Gyires-Tóth Bálint
  - Támogató szakértő: Járási István

- Jelenlegi megoldás:
  - Terméknevek kategorizálása TF-IDF módszerrel
  - CPU alapon
  - Vezető kutató, fejlesztő: Járási István

- Együttműködő szervezetek: NAV, NAV MIMCS, MNB, BME, SZTE, KSH

A mélytanulásról

- A gépi tanulás egyik fajtája.
- Elsődleges előny: jól skálázható, univerzális.
- Agyi neuronok ihlették, közöttük lévő kapcsolatok hálója
- Terminusok kölcsönzése:
  - Aktiváció
  - Rétegek: akár több száz rejtett
- Réteg típusok:
  - Teljes összeköttetésű
  - Rekurrens
  - Konvolúciós
- Bemenet →reprezentáció → kimenet



Mélytanulás és NLP

- Mélytanulás alapú NLP
  - Sok adatnál jobb teljesítmény, mint a hagyományos NLP módszerekkel
- Transformer alapú modellek
  - Teljes összeköttetés alapú
  - encoder-decoder alapú: szavak → vektorok → szavak
  - Input: szövegben random maszkolunk szavakat
  - Output: a hiányzó szó
- HuggingFace infrastruktúra (https://huggingface.co/)
  - Open source Python package

Putz Orsolya
putz.orsolya@tmit.bme.hu

A kutatás célja

MI eszközök használata az adatvezérelt döntéstámogatásban.

Rövidtáv

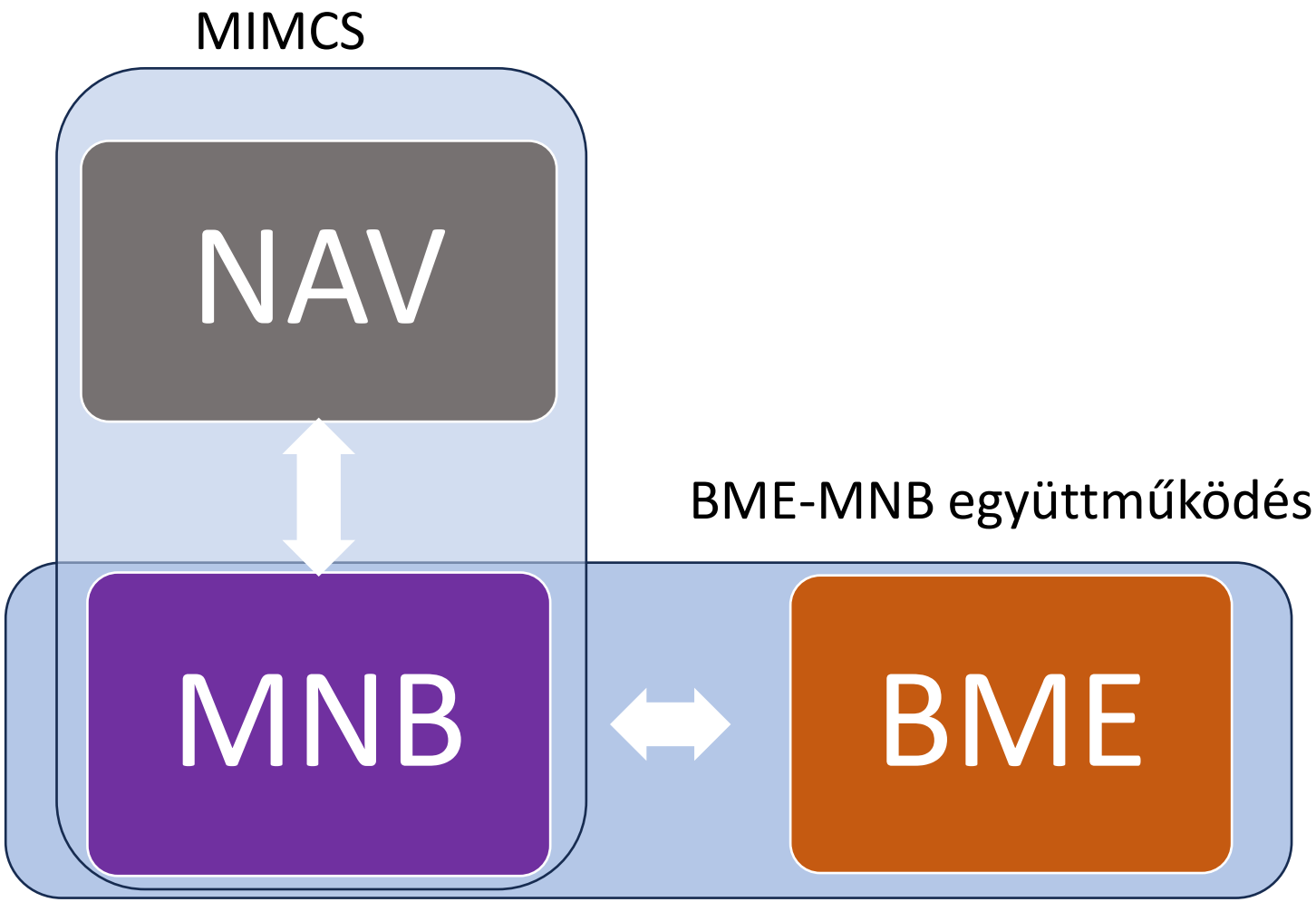
- OPG-s terméknevek besorolása termékcsoportokba
- A jelenlegi megoldás mellett a legújabb AI módszerek kipróbálása, egy robosztus, skálázható módszer fejlesztése

Középtáv

- OPG adatok automatizált feldolgozása és modellezése MI módszerekkel
- Adatból információ ember és gép közös munkájával

Putz Orsolya
putz.orsolya@tmit.bme.hu

Módszertan

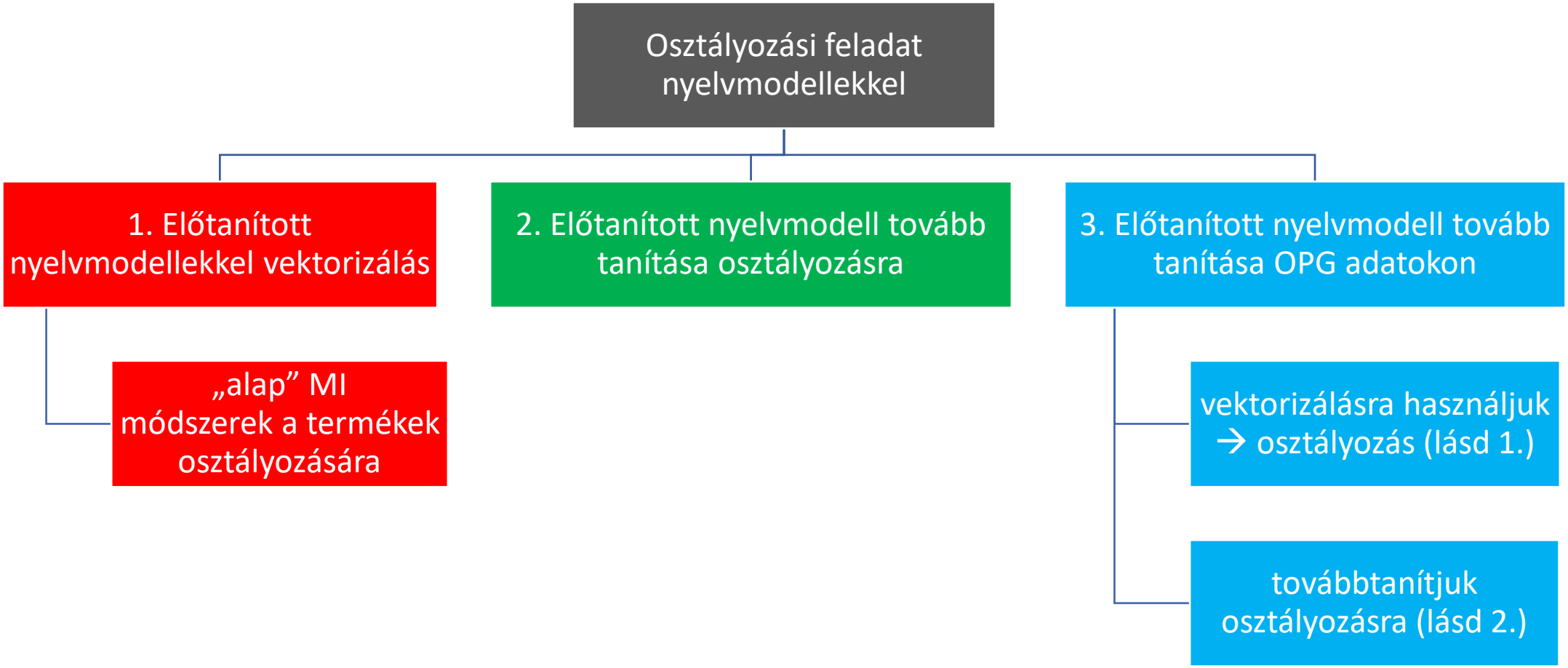


Munkafázisok

Az adatról

- Jellemzők
  - Első export:
    - 964.427 egyedi terméknév
    - Ebből vámtarifa számmal (VTSZ) rendelkezik: 88.276 rekord.
  - Második export:
    - 444.738 egyedi terméknév,
    - ebből 52.981-hez van VTSZ.
  - Korlátozottan hozzáférhető
- Adatgenerálás
  - Szabály alapon
  - Mélytanulásos alapon
    - ChatGPT-vel

Az MI szerepe: termékkategóriába sorolás mélytanulás alapú módszerekkel

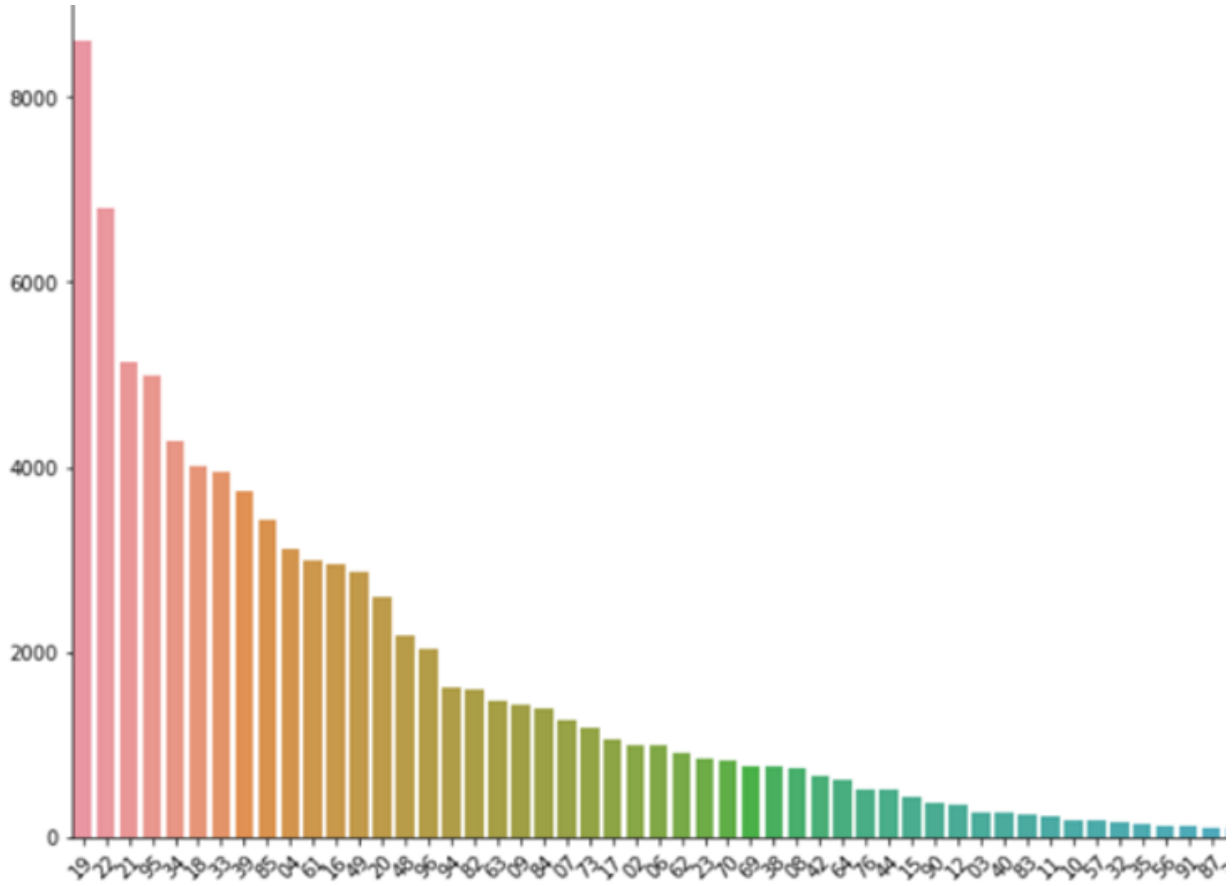


Putz Orsolya
putz.orsolya@tmit.bme.hu

Modellezési lépések

- Adattisztítás
  - üres cellák elhagyása
  - dirty vs clean adat
- Szétválasztjuk az adatot tanító és teszt adatbázisra.
- Mintavételi lehetőségek:
  - A termékek eloszlását figyelembe vesszük
  - Nem vesszük figyelembe = egyenletes mintavételezés
- Tanító-teszt adatbázis aránya: 70%-30% vagy 80%-20%.

elő-feldolgozás | tanítás | predikció | kiértékelés | hiperparaméter-beállítások | predikció



Putz Orsolya
putz.orsolya@tmit.bme.hu

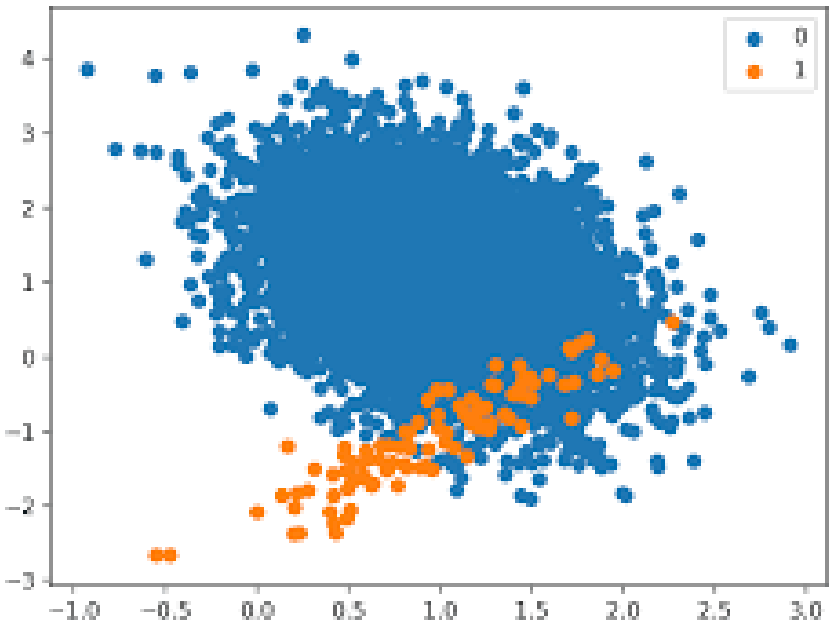
- Kétféle tanítás:
  - Előtanított nyelvi modell tovább tanítása osztályozásra (downstream task)
    - A nyelvmodellt termék-kategória párokon tanítjuk.
  - Előtanított nyelvi modell tovább tanítása OPG adatokon („saját” nyelvmodell)
    - A nyelvmodellt termékneveken tanítjuk.

- Prediktálás a teszt adatra.
- Feladat: olyan termékeket kategorizáljon a modell, amelyeket még nem látott.
- Mi tudjuk a „helyes választ”.

- A modell teljesítményének kiértékelése az alapján, hogy jó kategóriába sorolta-e a terméket.
- Mutatók:
  - Precision: tényleges pozitív / (tényleges pozitív + fals pozitív)
  - Recall: tényleges pozitív / (valódi pozitív + fals negatív)
  - Accuracy: (tényleges pozitív + tényleges negatív) / összes predikció
  - F1-score: a precision és a recall harmonikus átlaga

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

- Eredmények átlagolása: a nem egyenletes eloszlású osztályok kezelése céljából
  - Makro átlag: aggregátumokkal vagy összegekkel foglalkozunk, a mutatókat egészében vizsgáljuk.
  - Súlyozott átlag: figyelembe vesszük a gyakoriságot.



Putz Orsolya
putz.orsolya@tmit.bme.hu

- A modell teljesítményének javítása a legoptimálisabb paraméterek megtalálásával.
- Fontosabb hiperparaméterek:
  - Batch size: mennyi tanuló adatot adunk be a modell frissítése előtt.
  - Number of epochs: hányszor menjünk végig a tanító adatokon.
  - Learning rate: súlyokat állítunk a neuronok közötti kapcsolat kialakítására.

Putz Orsolya
putz.orsolya@tmit.bme.hu

- Prediktálás az ismeretlen adatokra.
- Feladat: olyan termékeket kategorizáljon a modell, amelyeket még nem látott.
- Nem tudjuk a „helyes választ”.
- Azt tudjuk, hogy a modellünk 80%-ban jól dolgozik.

Putz Orsolya
putz.orsolya@tmit.bme.hu

Modellezés 3 szinten

1. Termékcsoportok elemszáma alapján

- Kizárólag azokat a termékcsoportokat vontuk be a tanításba, amelyek 1000 elemet tartalmaznak

2. A vámtarifaszám szintjei alapján

- A VTSZ_4 (első négy számjegy), két szintjét kellett megtanulnia a modellnek

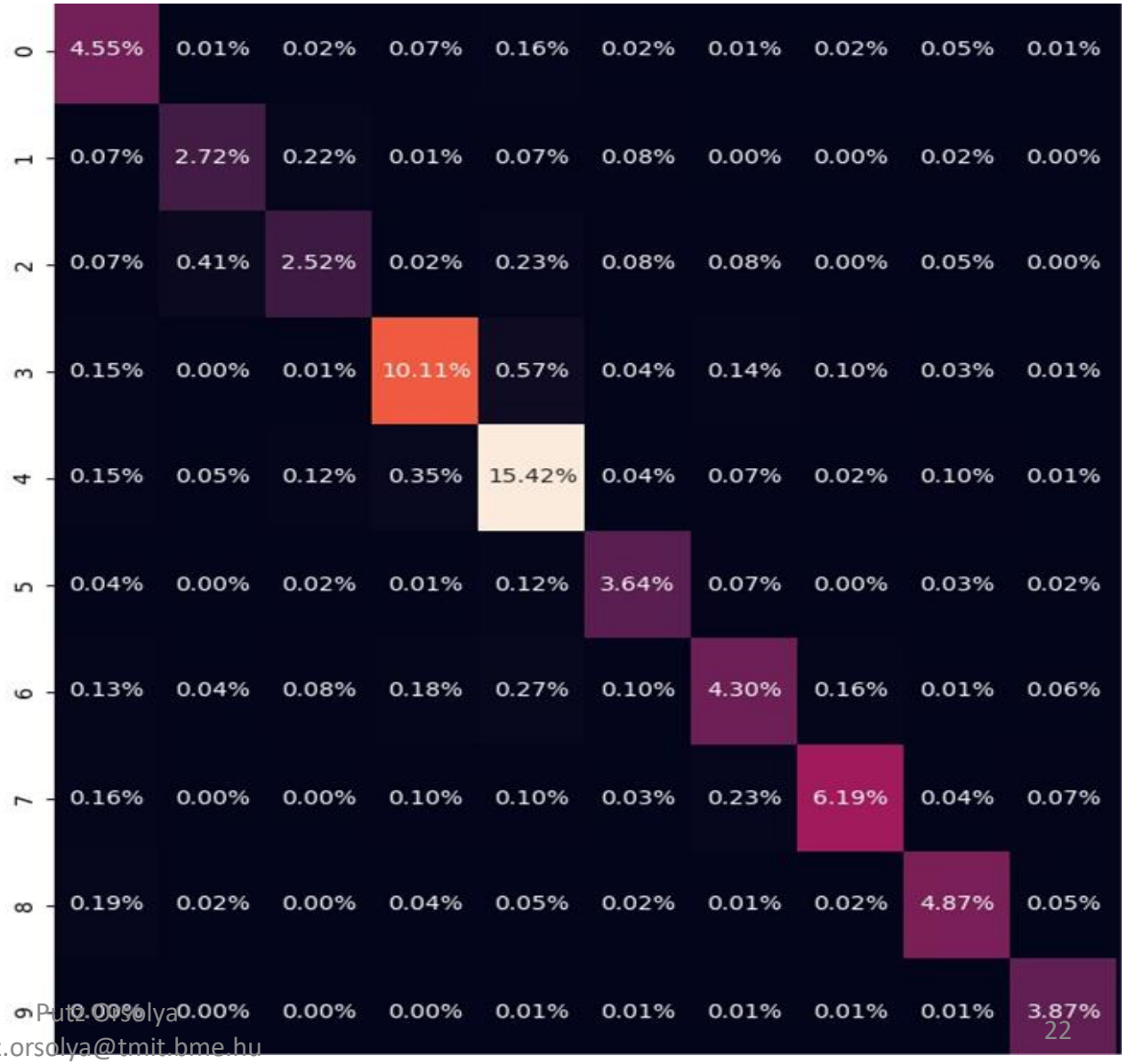
3. A vámtarifaszám szintjei alapján

- A VTSZ_2 (első két számjegy), egy szintjét kellett megtanulnia a modellnek

1. lépés eredményei: min. 1000 terméknév/ kateg.

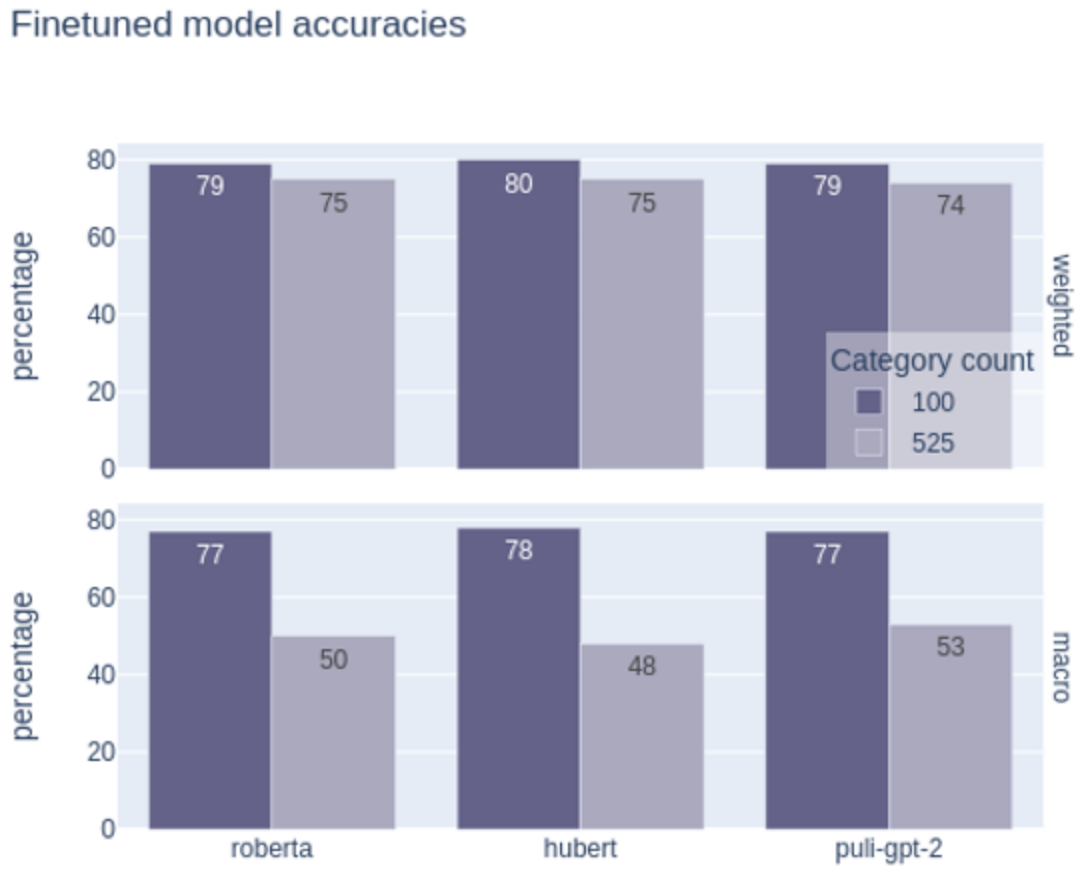
10 leggyakoribb kategória

	precision	recall	f1-score
1	0.81	0.92	0.86
2	0.83	0.85	0.84
3	0.84	0.72	0.78
4	0.92	0.89	0.90
5	0.88	0.93	0.91
6	0.89	0.91	0.90
7	0.86	0.80	0.83
8	0.94	0.88	0.91
9	0.91	0.91	0.91
10	0.86	0.88	0.87
acc.			0.88
m. avg	0.86	0.86	0.86
w. avg	0.88	0.88	0.88



2. lépés eredményei: VTSZ 2 szintje

- 100 kategória
  - min. 130 termékkel
- 525 kategória
  - min. 2 termékkel



2. lépés eredményei: VTSZ 2 szintje

- 525 kategória, min. 2 termékkel → kategóriák összevonása

→ 213 kategória

accuracy			0.72
macro avg	0.65	0.59	0.59
weighted avg	0.71	0.72	0.70

3. lépés eredményei: VTSZ 1 szintje

- folyamatban

Következő lépések

- Legújabb nyelvmodellek bevezetése
  - pl. Meta AI Llama2
  - → hozhat-e további javulást a még nagyobb modell?
- Szintetikus adatgenerálása LLM-el (pl. ChatGPT)
- Termékkategóriákat tartalmazó idősoros

adatok elemzése

- További adatkörök bevonása az elemzésbe

Köszönöm a figyelmet

putz.orsolya@tmit.bme.hu