Magyar nyelvű Szentiment Analízis Projekt

Név

2025. április 24.

Tartalomjegyzék

1. Projekt Áttekintés

Ez a projekt célja egy magyar nyelvű szentiment analízis modell fejlesztése Pythonban, amely a HuSST adatkészletet használja. A modell felé elvárás, hogy képes legyen szövegeket negatív, semleges és pozitív kategóriákba sorolni.

2. Módszertan

A cél megvalósításához a huBERT betanított neurális hálót fogom felhasználni alapmodellként. Az előre betanított neurális háló nagyon jó kiindulási alapként szolgál, mivel magyar nyelvű adatokon tanították tehát általános magyar nyelvtudással rendelkezik. Képes a szövegek értelmezésére és feldolgozására, viszont általánosságban elmondható, hogy ezeket az alapmodelleket további tanítással kell kiegészíteni ha specifikusan egy bizonyos célra szeretnénk használni a tudását.

Jelen feladatban a HuSST adathalmazzal fogok további tanítást végezni a modellen. A HuSST mint korábban említsre került, magyar nyelvű kijelentéseket tartalmaz és az azokhoz tartozó címkét. A címke lehet negatív, semleges, vagy pozitív. Ezek alapján kerül besorolásra az adott szöveg.

3. Dataset

A bevezetőben ismertetett két forrást fogom használni a projekt megvalósításához.

- huBERT base model (Hungarian Universal Bidirectional Encoder Representations from Transformers)
- HuSST dataset (Hungarian Stanford Sentiment Treebank)

3.1. huBERT bemutatása

A huBERT egy magyar nyelvű, transzformátor alapú nyelvi modell, amelyet a SZTAKI fejlesztett ki. A modell a BERT architektúrát követi, és kifejezetten a magyar nyelv sajátosságainak kezelésére optimalizálták. A tanítást az

úgynevezett *Common Crawl* adatbázis magyar nyelvű részén végezték szűrések és deduplikációk után, valamint a magyar Wikipedia alapján. A modell 111 millió paraméterrel rendelkezik.

3.2. A huBERT alkalmazási lehetőségei

A huBERT modellt különféle magyar nyelvű NLP feladatokhoz használhatjuk:

- Szövegosztályozás
- Névvelentismerés (NER)
- Szövegrészletezés (chunking)
- Kérdésmegválaszolás
- Szöveggenerálás

4. Implementáció

A modell Pythonban készül a következő könyvtárakkal:

- `torch` és `torch.nn`: A neurális hálók megvalósításához és tensor műveletekhez
- `torch.optim`: Optimalizálási algoritmusok (pl. Adam, AdamW, SGD)
- `torch.utils.data`: Adatbetöltés és előfeldolgozás
- `sklearn.metrics`: Osztályozási metrikák kiértékelése
- `transformers`: Előtanított nyelvi modellek és tokenizálók
- `datasets`: Nagy nyelvi adathalmazok kezelése
- `pandas`: Adatkezelés és -elemzés
- `numpy`: Numerikus számítások
- `tqdm`: Progress bar a betanítás során
- `os`: Operációs rendszer szintű műveletek (pl. fájlkezelés)

5. Források

A dokumentumot az alább felsorolt források segítségével készítettem el.

Hivatkozások

[1] SZTAKI-HLT. (2022). *hubert-base-cc*. Hugging Face.
`https://huggingface.co/SZTAKI-HLT/hubert-base-cc`

[2] NYTK. (2022). *HuSST Dataset*. Hugging Face.
`https://huggingface.co/datasets/NYTK/HuSST`

[3] SZTAKI-HLT. (2022). *huBERT - Hungarian BERT Model*. BME-HLT.
`https://hlt.bme.hu/hu/resources/hubert`

[4] Orosz György. (2023). *Awesome Hungarian NLP Resources*. GitBook.
`https://oroszgy.gitbook.io/awesome-hungarian-nlp-resources`

[5] Orosz György. (2023). *Awesome Hungarian NLP*. GitHub.
`https://github.com/oroszgy/awesome-hungarian-nlp`

[6] Laki László J., Yang Zijian Győző. (2022). *huBERT - Hungarian BERT*. Acta Universitatis Óbuda.
`https://acta.uni-obuda.hu/Laki_Yang_134.pdf`