Data kihívás (2025, nyár):
web scraping + AI-os érzelemelemzés

Poet.hu oldal verseinek érzelemelemzése BERT modellel

készítette:
Kunsági Zsolt

A **Poet.hu** (poet.hu) Magyarország egyik legnagyobb és leglátogatottabb amatőr verses oldala, több mint 38 000 regisztrált felhasználóval és naponta átlagosan 10 000 egyedi látogatóval (2016. májusi adatok).

Poet.hu tapasztalat: költői álnéven én is küldtem be verseket. Kedves, támogató közösség az oldalon.

Projekt cél: BERT modellel érzelemelemzés. Miért nem ChatGPT? Ingyenes verzióban hamar OFF-olt. Érdekelt egy olyan megoldás megvalósítása, ahol nem használok külső szolgáltatást. Pl: nagy mennyiségű adat feldolgozása esetén hasznos.

Az AI program el tudja dönteni, hogy egy vers Pozitív vagy Negatív hangulatú, kategóriájú?

Web scraping: Google Notebooks-ban oldottam meg. Tapasztalat: ha nem Google felől érkezett a nagy mennyiségű adatletöltés, akkor a Poet.hu szerver egy idő után időtúllépés zárja a kapcsolatot.
Colab link: https://colab.research.google.com/drive/1B0-N8vKak0iOAYj9VWlgxewLU7bQ1pu4?usp=sharing

Projekt fájlok, GitHub link: https://github.com/minorharpman/erzelemelemzes/tree/main

Poet.hu Érzelmek oldalán (poet.hu/Erzelmek/) a
Pozitív kategóriájú versek: **Szeretet, Boldogság, Bizalom, Mosoly**
Negatív versek: **Félelem, Sírás, Fájdalom, Szenvedés** kategóriában találhatók.

A Tanító és Valid elemzéshez verseket: **vers_adatok.xlsx** fájlba lett mentve.
Ellenőrzéshez (Teszt) a **ellenorzo_vers_adatok.xlsx** fájlt használtam.

Az Excel fájlok oszlopai: Típus, Szerző, Cím, Vers, Link.
A Tanító és Valid verseknél (vers_adatok.xlsx) a Pozitív besorolású verseket átnéztem, a köztük a negatív hangulatúakat OFF jelzővel láttam el. (további szűrések: kizárások, duplikációk)

Pozitív versek száma: **88**, Negatív versek: **90** (itt töröltem **2**-t a DataFrame-ből.)
Train halmaz mérete: **140**, Valid halmaz mérete: **36** (80%-20% -os elosztás)

Érzelemelemzés BERT-alapú modell segítségével (SZTAKI-HLT/hubert-base-cc)
Beállítások: Epoch =**6** (Kísérleteztem: 5,10, 7 beállításokkal)
Learning rate = **5e-6**

Train ábra:

Epoch	Training Loss	Validation Loss	Accuracy
1	0.6981	0.671959	0.527778
2	0.6648	0.615367	0.805556
3	0.5064	0.557467	0.861111
4	0.4265	0.507214	0.833333
5	0.3901	0.479909	0.833333
6	0.3859	0.469884	0.833333

Ellenőrzés, hogyan teljesít a modell
Az ellenorzo_vers_adatok.csv segítségével, 221 db verssel folyt a tesztelés
Biztonsági ellenőrzés a Tanító-Valid versek és az Ellenőrző (Teszt) versek halmazán nem lehet azonos vers.

Predikció:
Tároltam az eredeti Pozitív/Negatív típust. A tanított Bert modellt lefuttattam versenként az Ellenőrző versek halmazán, rögzítettem a Predikció (Pozitív/Negatív) értékét. Az eredményt a **Kimenet versek** (kimenet_versek.xlsx) Excelben tároltam.
https://github.com/minorharpman/erzelemelemzes/tree/main

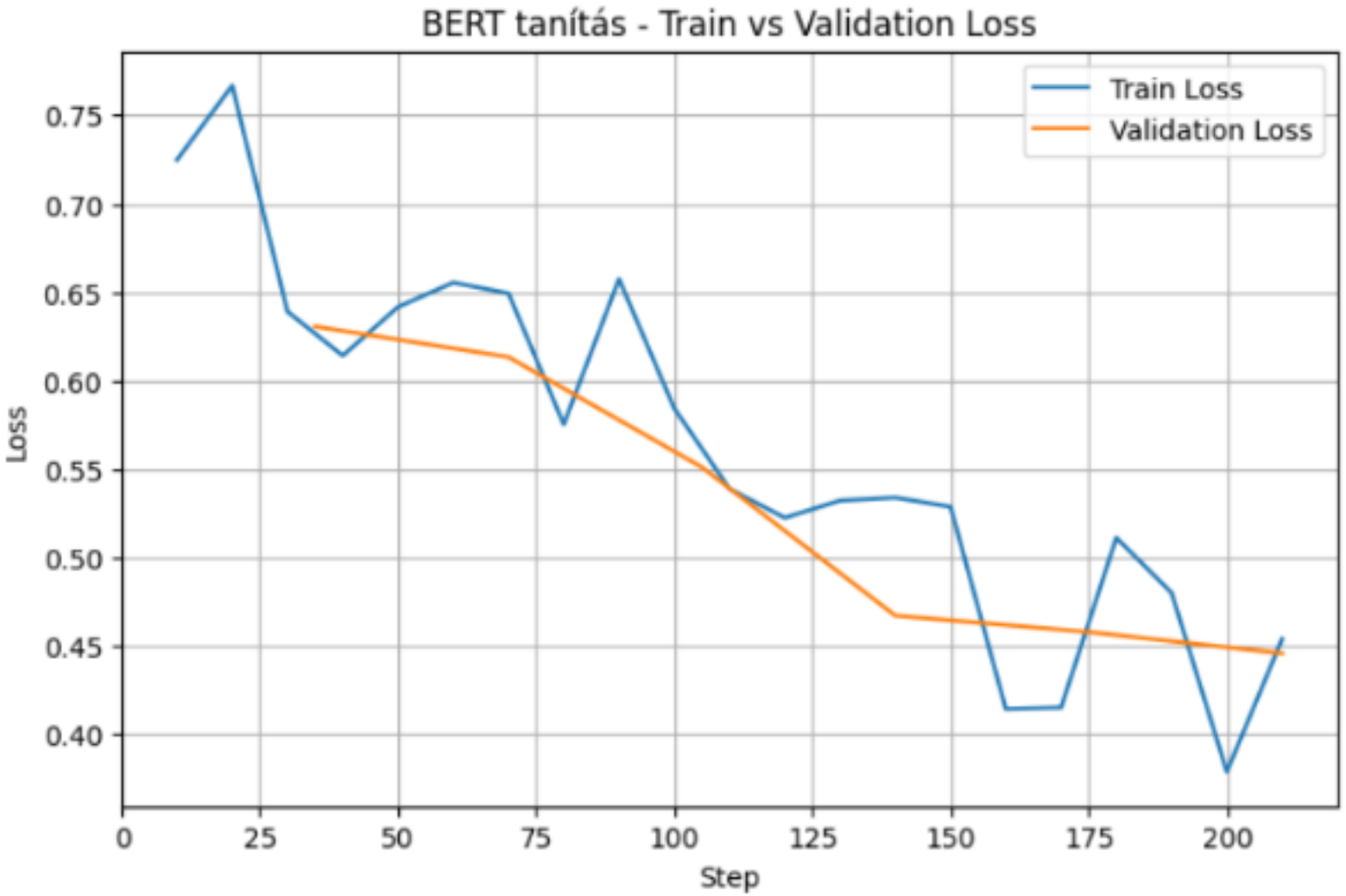
A Kimenet excel (Kimenet DataFrame) szerint: **76,47% a pontosság** az éles adatokon.
A tanított Bert modell trainer.evalute() eval_accuracy: **83.33%** volt pontosság.

Tapasztalat, ötletek:
Fontos, hogy a Train-Valid halmazba megfelelően kategorizált adatok kerüljenek.
Ki legyen egyensúlyozva a kategóriák.
Web scraping-nél a blokkolás kivédése
Más modellt is kipróbálni.
Verseknél soha nem lesz tökéletes a Predikció, a „költői boldogságban” írt verset nehéz a gépnek kategorizálni. (lásd a Kimenet versek azon verseit, ahol a Valódi oszlop értéke nem egyezik a Predikció oszlop értékével.

BERT modell tanítás görbéje

A modell jól tanul, a **Train Loss** fokozatosan csökken**.**

A **Validation Loss** görbe is fokozatosan csökken, nem kezd emelkedni. Nem tanulta túl az adatokat.



További ötletek:

A magyar nyelv nehéz, árnyalt kifejezésekre képes nyelv.

További tapasztalat szerzés más területeken.

Pl: Cikkek kategorizálása, Kommentek elemzése, Bulling, Indirekt sértő kifejezések szűrése,

Fenyegető üzenetek szűrése ...

Köszönöm a figyelmet!