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

Poet.hu oldal verseinek érzelemelemzése BERT modellel

készítette:
Kunsági Zsolt
Infodok

Poet.hu

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. Érdekelt egy olyan megoldás megvalósítása, ahol nem használok külső szolgáltatást (ChatGPT). Pl: nagy mennyiségű adat feldolgozása esetén hasznos. (Többszörfuttattam le a kódot. Nem fogyasztottam tokent.)

Kezdeti célkitűzés:

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

Web scraping: Google Colab Notebooks-ban oldottam meg a Beautiful Soup könyvtárral.
Weboldal linkek, CSS osztálynevek szerint találtam meg az adatokat.
Tapasztalat: ha nem Google felől érkezett a nagy mennyiségű adatletöltés, akkor a Poet.hu szerver egy idő után időtúllépés zárja a kapcsolatot.

Web scraping: Poet.hu Érzelmek oldalán (poet.hu/Erzelmek/) a

Pozitív kategóriájú versek: **Szeretet, Boldogság, Bizalom, Mosoly**

Negatív versek: **Félelem, Sírás, Fájdalom, Szenvedés** kategóriában találhatók.

(Éles szétválasztás, az elégikus, melankolikus hangulatúakat, kategóriába tartozókat nem vettem bele.)

Modell készítéshez és a Teszteléshez 3db Excelt készítettem (A, B, C) web scraping módszerrel.
(**A**:189, **B**: 219, **C**: 240 db verset tartalmazott.)

Az Excel fájlok oszlopai: **Típus (pozitív/negatív)** , Szerző, Cím, **Vers**, Link.

Adatok vizsgálata: A verseket átnéztem (átfutottam). Néhányat OFF jelzővel láttam el.
Duplikációkra figyeltem, a tanító, validációs, teszt adatoknál ne legyen azonos vers.

Kiegyensúlyozás: A Modell tanítására szánt Exceleket (pl. A+B) előbb kiegyensúlyoztam, egyenlő Pozitív és Negatív hangulatú vers legyen a táblában.
A Tanító és Validációs arány: 70%-30%
Pl. A+B (A és B Excelek egyesítésével) a Train halmaz mérete: 261, Validációs halmaz mérete: 113

BERT modell

(Bidirectional Encoder Representations from Transformers)
Google AI, 2018

BERT modell a **Transformer neurális hálózati struktúrán alapul**, amely jól kezeli
a szöveges adatokban lévő hosszú összefüggéseket.

Bidirectional (kétirányú) kontextus: A korábbi modell csak balról jobbra olvasta a szöveget,
míg BERT **egyszerre veszi figyelembe a szavak bal és jobb kontextusát**.

A **BERT modell előre betanított (pre-trained)** modell. Már nagy mennyiségű szövegen megtanították neki
a nyelv általános mintáit és szabályszerűségeit.

A BERT modellnek a SZTAKI-HLT/hubert-base-cc változatát használtam (magyar nyelvre optimalizált,
pre-trained BERT-típusú modell.)

Érzelemelemzés BERT-alapú modell segítségével

(SZTAKI-HLT/hubert-base-cc)

Fine-tuning: Előre betanított modell egy adott feladatra hangolása

Beállítások:

Epoch =9 (Kísérleteztem: 5,6,7,8,9,10 beállításokkal)

1 **epoch** azt jelenti, a modell **egyszer teljesen végigmegy az egész tanító adathalmazon**.

Learning rate = 1e-5

A learning rate azt határozza meg, hogy a modell mekkora lépésekben változtatja a súlyait az optimalizálás során.

Batch size: egyszerre hány tanítópéldát adunk be a neurális hálónak, mielőtt frissíti a súlyait.

Az adatokkal (versekkel) tanítás (Fine-tuning) :

[2132…

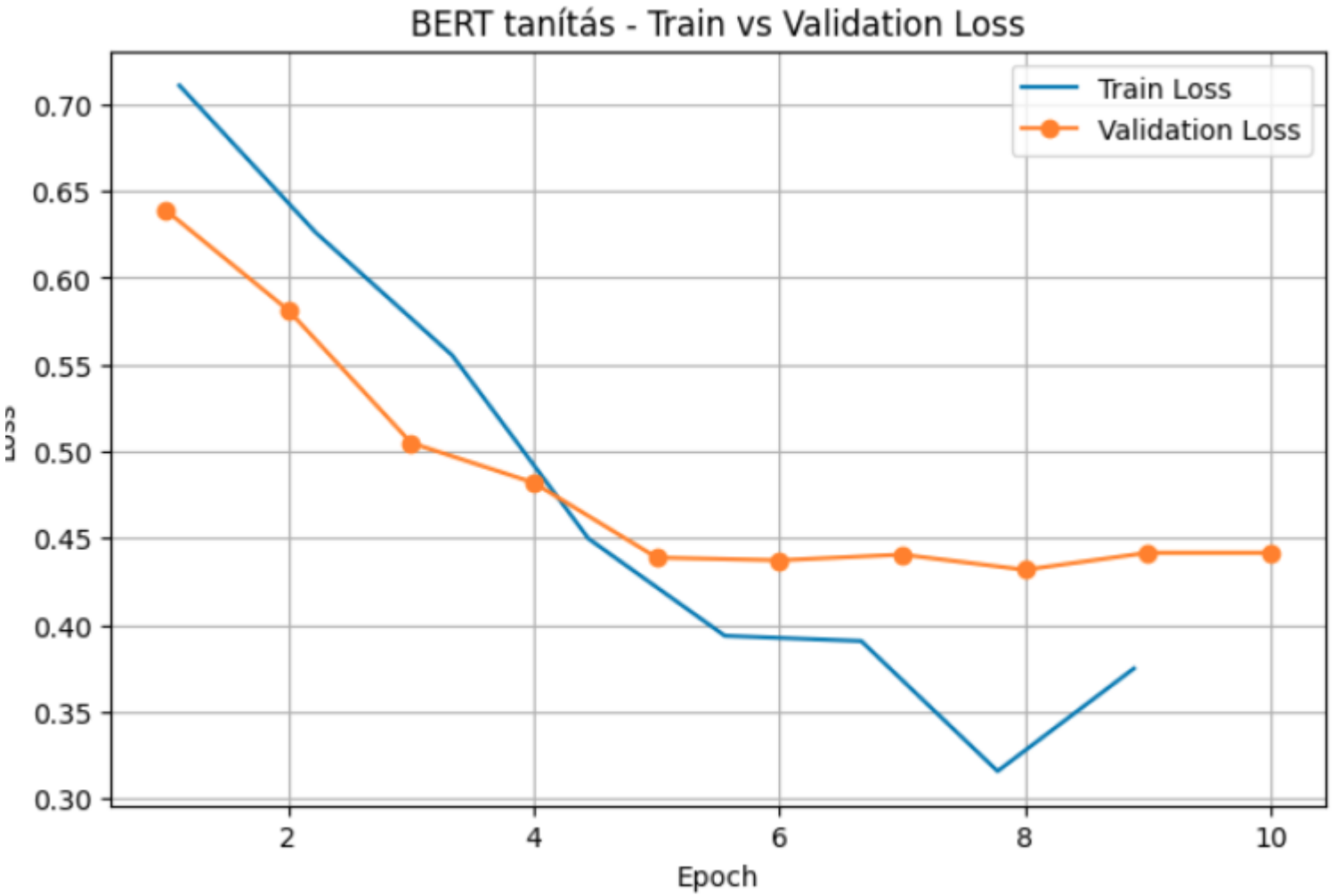
```
# ===== 9. Modell tanítása =====
trainer.train()
```

[81/81 01:44, Epoch 9/9]

Epoch	Training Loss	Validation Loss	Accuracy	Auc	Precision	Recall	F1
1	No log	0.639017	0.681416	0.719925	0.616279	0.946429	0.746479
2	0.710900	0.581263	0.805310	0.822995	0.765625	0.875000	0.816667
3	0.626100	0.504878	0.805310	0.859023	0.783333	0.839286	0.810345
4	0.555400	0.481884	0.743363	0.880326	0.675325	0.928571	0.781955
5	0.449700	0.438789	0.787611	0.884712	0.750000	0.857143	0.800000
6	0.393700	0.437225	0.805310	0.895050	0.757576	0.892857	0.819672
7	0.390700	0.440525	0.796460	0.897556	0.746269	0.892857	0.813008
8	0.315700	0.431743	0.814159	0.896617	0.769231	0.892857	0.826446
9	0.374900	0.441562	0.787611	0.898183	0.735294	0.892857	0.806452

[2133…

BERT modell tanítás görbéje (9 epoch)
A modell jól tanul, a **Train Loss** fokozatosan csökken**.**
A **Validation Loss** görbe is fokozatosan csökken, nem kezd emelkedni. Nem tanulta túl az adatokat.



Tanítás értékelése

A+B Excelekkel tanítás
Accuracy (Pontosság): 79%
(Hányat talált el a modell a tanítás során)
Tesztelés pontosság C Excel adataival: 77%
AUC: 0.89
(Modell mennyire tudja megkülönböztetni a pozitív és negatív osztályokat. 89% esély)

A+C Excelekkel tanítás
Accuracy (Pontosság): 77%
Tesztelés pontosság B Excel adataival: 78%
AUC: 0.85

B+C Excelekkel tanítás
Accuracy (Pontosság): 78%
Tesztelés pontosság A Excel adataival: 86%
AUC: 0.86

A modell tévesztésének vizsgálata

Tanító-Valid Excel	Teszt Excel	Valódi	Predikció	Tévedés
A+B	C	neg	poz	8
A+B	C	poz	neg	22
A+C	B	neg	poz	21
A+C	B	poz	neg	10
B+C	A	neg	poz	11
B+C	A	poz	neg	10

Nincs valamilyen határozott irányba, pl. Negatívból Pozitívba tévedés. Egyenlően téved a modell Negatív vagy Pozitív fele.

A predikciók szubjektív vizsgálata

Érthető, hogy a BERT modell téveszt, hogy a vers pozitív vagy negatív hangulatú.
Ez már irodalmi verselemzés szintje, annak megállapítása, hogy pl. a költő negatív képek használata ellenére is derűs hangulatú verset írt.

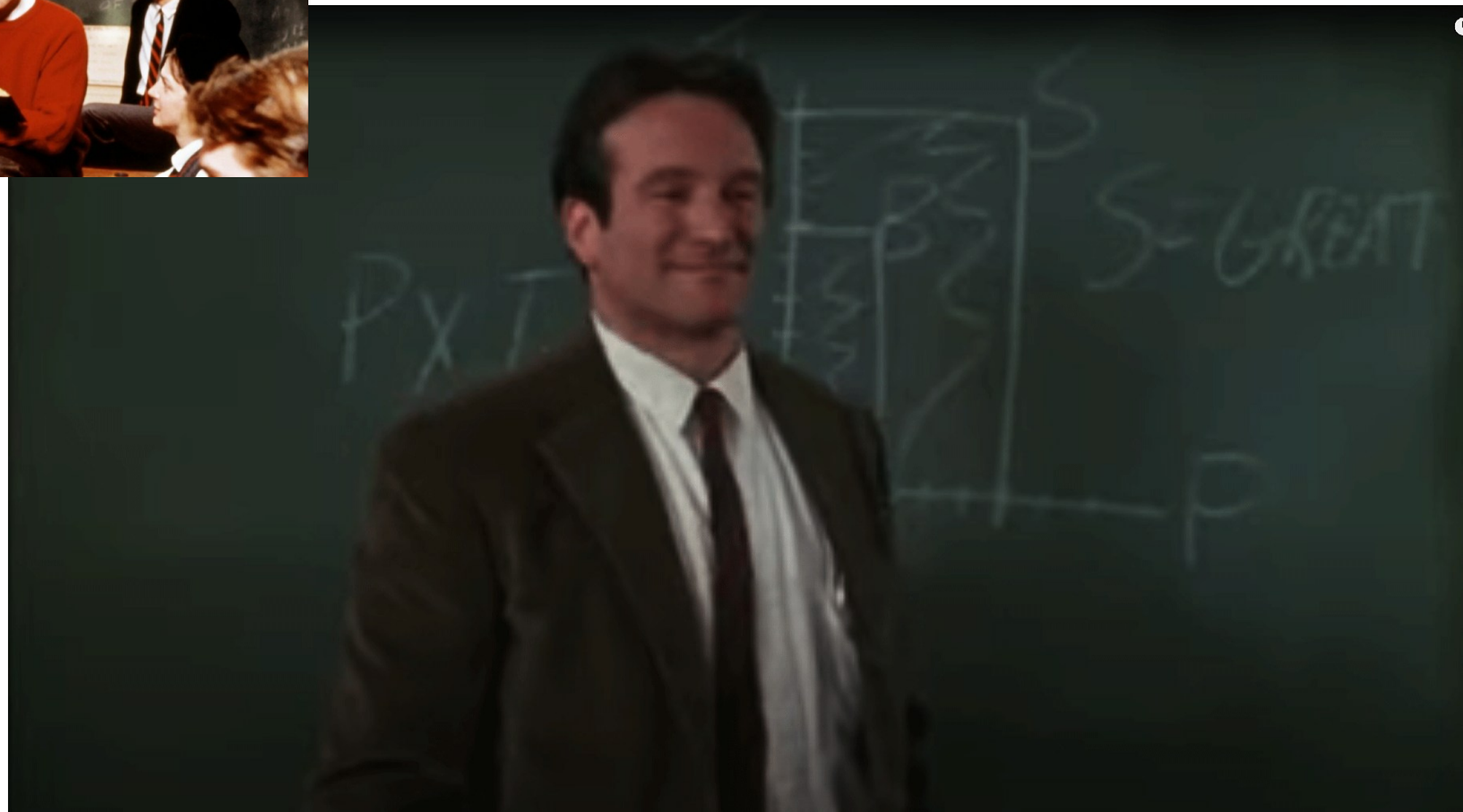
Siralmas énnéköm...	Százmilliós kégli kéne? Nem reális, csak álom. Az árak égbe repültek, ezért nem találom. Azért mégis érdekelne, hogy "Vajjon s mikor leszön jó Budában lakásom!" Vidéki lét. Levegő ép. Bár bezárt a gyár - azt látom. Nem hiszem el, hogy nincs, ki felel: "Vajjon s mikor leszön jó Budában lakásom!" Bejárni jó, csak időrabló, de csökken a bérlettel károm. Irigykedve kérdem e röpke létben: "Vajjon s mikor leszön jó Budában lakásom!" Kánikula. Melegrekord. Síkságon sültem a nyáron. Elképzelem, hogy a fővárosban élek, s a hegyeket látom. "Vajjon s mikor leszön jó Budában lakásom!"	https://poet.hu/vers/381061	poz	neg
Hangod	Ha te sem hallod a saját hangod, Más hogy hallaná meg, Csak tűröd némán, oktalan, Mit más mond, s semmit nem Hallatsz magadból, Légy hát, ki vagy, Ki lehetnél, lelj rá, S önmagad képmása Sem leszel többé, Tedd vagy ne, De dönts csak te, Mert jogod van magadhoz, Jogod van annak lenni, Ki vagy, ki lehetnél, kinek születtél, Kit nem nyom el más, Sem az, mit magáról hisz, Vagy nem hisz Vagy nem tudja már, Ki lehetne s mit hisz, Mert önmagát sosem Látta, csak a szörnyet, mi lehetne, Pedig csodásabb volt, mint a Világ hét csodája, Mert önmaga felért mindjével Csak önmagad légy, Tedd, mi zsigerből jön, S meglásd, hibázhatsz is, Ilyen vagy, mindnyájan ilyenek volnánk Ebben a valóra vált mesében Hibázunk, tanulunk, fejlődünk, Önmagunkból csak önmagunk Lehetünk, Fényesebb vagy, mint hinnéd, Gyönyörűbb bármely mesénél, Csak higgy magadban, S abban, hogy bármikor Lehetsz önmagad, Mert ki nem fogad el olyannak, Milyen vagy, Az olyannak sem fogadna el, Amilyenné formálna, Amilyen azt gondolja, hogy vagy, Hallasd	https://poet.hu/vers/380270	poz	neg

Megállapítás, összefoglalás, :

A BERT modellel (SZTAKI-HLT/hubert-base-cc) 78% pontossággal sikerült az amatőr költők verseiről eldönteni, hogy pozitív vagy negatív hangulatú.

A Poet.hu oldalról a pozitív és a negatív kategóriájú verseket 3 db Excelbe mentettem.
Az adatokat megvizsgáltam, kizártam néhány verset, töröltem a duplikációkat.
2db Excel felhasználásával tanítottam a modellt. A 3. Excellel teszteltem
a működést. Minden variációt kipróbáltam (A+B, teszt: C, A+C, teszt: B, B+C, teszt: C), közel azonos eredményekre jutottam.

Holt költők társasága



További ötletek:

További tapasztalatszerzés más területeken.
Pl: Cikkek kategorizálása, Kommentek elemzése, Bulling, Indirekt sértő kifejezések szűrése,
Fenyegető üzenetek szűrése ...

Köszönöm a figyelmet!