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. (Többször 20-30x futtattam 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.)

A Modell készítéshez (Tanító és Validációs adatok) verseket: **vers_adatok.xlsx** fájlba lett mentve.
Ellenőrzéshez (Teszt) a **ellenorzo_vers_adatok.xlsx** fájlt használtam.

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

Adattisztítás: A Modell készítéshez használt adatoknál (vers_adatok.xlsx) a Pozitív besorolású verseket átnéztem, a köztük a negatív hangulatúakat OFF jelzővel láttam el (kivettem). (további szűrések: kizárások, duplikációk)

Kiegyensúlyozás: Pozitív versek száma: **88**, Negatív versek: **88** (DataFrame műveletek, hogy egyforma legyen)
Train halmaz mérete: **140**, Validációs halmaz mérete: **36** (80%-20% -os elosztás)

BERT modell

(Bidirectional Encoder Representations from Transformers)

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 =8 (Kísérleteztem: 5,6,7,8,9 beállításokkal)
1 **epoch** azt jelenti, a modell **egyszer teljesen végigmegy az egész tanító adathalmazon**.
Learning rate = 5e-6
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.

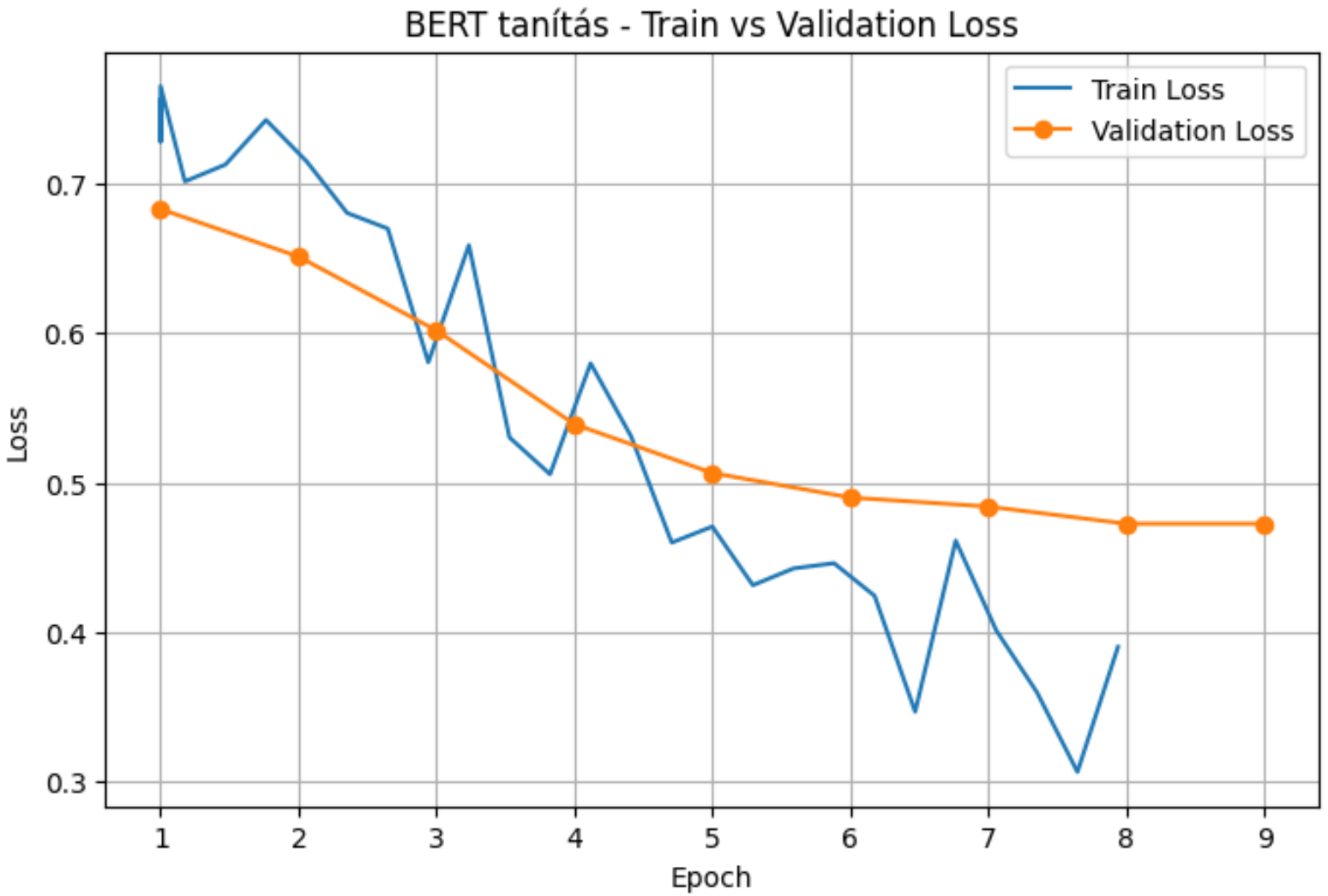
Az adatokkal (versekkel) tanítás
(Fine-tuning) :

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

[272/272 00:45, Epoch 8/8]

Epoch	Training Loss	Validation Loss	Accuracy	Auc	Precision	Recall	F1
1	0.765000	0.682809	0.571429	0.535948	1.000000	0.117647	0.210526
2	0.742400	0.651052	0.742857	0.542484	0.785714	0.647059	0.709677
3	0.580400	0.601649	0.742857	0.712418	0.750000	0.705882	0.727273
4	0.505600	0.539025	0.771429	0.810458	0.800000	0.705882	0.750000
5	0.470600	0.506417	0.771429	0.833333	0.800000	0.705882	0.750000
6	0.446100	0.490003	0.771429	0.839869	0.800000	0.705882	0.750000
7	0.461200	0.483972	0.800000	0.843137	0.857143	0.705882	0.774194
8	0.390200	0.472495	0.771429	0.843137	0.800000	0.705882	0.750000

BERT modell tanítás görbéje (8 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.



A finomhangolt (**Fine-tuning)** BERT modell, tanítás **értékelése**

Accuracy (Pontosság): 77.14%
(Hányat talált el a modell a tanítás során)
AUC: 0.84
(Modell mennyire tudja megkülönböztetni a pozitív és negatív osztályokat. 84% esély)

Precision: 0.80
(Modell által pozitívnak jelzett példák 80%-a valóban pozitív volt.)
Recall: 0.7059
(Modell a valódi pozitív esetek 70,59% -t megtalálta.
F1: 0.7500
(Precision és a Recall harmonikus átlaga. Kiegyensúlyozatlan osztályoknál jobb összegző mutató, mint az Accuracy.)

(Jupyter notebook oldal HTML-ként, feltölt Google colab, majd ment HTML)

A finomhangolt (**Fine-tuning)** BERT modell, tanítás **értékelése**

Accuracy (Pontosság): 77.14%
(Hányat talált el a modell a tanítás során)
AUC: 0.84
(Modell mennyire tudja megkülönböztetni a pozitív és negatív osztályokat. 84% esély)

Precision: 0.80
(Modell által pozitívnak jelzett példák 80%-a valóban pozitív volt.)
Recall: 0.7059
(Modell a valódi pozitív esetek 70,59% -t megtalálta.
F1: 0.7500
(Precision és a Recall harmonikus átlaga. Kiegyensúlyozatlan osztályoknál jobb összegző mutató, mint az Accuracy.)

(Jupyter notebook oldal HTML-ként, feltölt Google colab, majd ment HTML)

Tesztelés, hogyan teljesít a modell

Az **ellenorzo_vers_adatok.csv** segítségével, 180 db verssel folyt a tesztelés

Adattisztítás: Biztonsági ellenőrzés a **Tanító** versek és az **Teszt** versek halmazán nem lehet azonos vers.

Teszt:
Itt is volt **Típus oszlop (pozitív/negatív).** A tanított Bert modellt lefuttattam versenként az Ellenőrző versek halmazán, rögzítettem a Predikció (Pozitív/Negatív) értékét.

77,63% a pontosság az éles adatokon.
(Előző slide: A tanításnál Bert modell **Accuracy**: **77.14%** volt).

Fordított vizsgálat

Korábban: A Modell készítéshez (Tanító és Validációs adatok) verseket: **vers_adatok.xlsx** fájlba lett mentve.
Ellenőrzéshez (Teszt) a **ellenorzo_vers_adatok.xlsx** fájlt használtam.

Pontosság: 77.14%
AUC: 0.8431
Precision: 0.8000
Recall: 0.7059
F1: 0.7500
sz. Pontosság: 77.63%

Most: A Modell készítéshez (Tanító és Validációs adatok) verseket: **ellenorzo_vers_adatok.xlsx** fájlba lett mentve.
Ellenőrzéshez (Teszt) a **vers_adatok.xlsx** fájlt használtam.
Az ellenorzo_vers_adatok.xlsx -t nem néztem át. Nem adattisztítás.

Pontosság: 76.19%
AUC: 0.8980
Precision: 0.8235
Recall: 0.6667
F1: 0.7368
sz. Pontosság: 73.33%

Saját verseken teszt

Vers: Makesz ma kezd. Kezében keksz. Mit kapot... Predikció: poz (pozitív) | Valódi: poz (pozitív)
Vers: Tudsz, strucc, a szobában ülni? Tudsz, s... Predikció: poz (pozitív) | Valódi: poz (pozitív)
Vers: Szeretem, ha süt rám a Nap. A tóból kifo... Predikció: poz (pozitív) | Valódi: poz (pozitív)
Vers: Kiflit enni volna jó! De nem volt nálam ... Predikció: poz (pozitív) | Valódi: poz (pozitív)
Vers: az ég nagyon kék a fű zöld és teli van p... Predikció: poz (pozitív) | Valódi: poz (pozitív)
Vers: Nem tudok focizni. Nem vesz be a csapat.... Predikció: poz (pozitív) | Valódi: neg (negatív)
Vers: Esik az eső a prolira. Megy a gyerek ovi... Predikció: poz (pozitív) | Valódi: neg (negatív)
Vers: Útra keltem megkeresni az Arany-hegyet. ... Predikció: poz (pozitív) | Valódi: neg (negatív)

A vicces, vidám hangulatúakat jól prediktálta.

Az elégikus, mélabús hangulatúakat is inkább a pozitív kategóriába sorolta.

További ötletek:

A magyar nyelv nehéz, árnyalt kifejezésekre képes nyelv.
További tapasztalatszerzés más területeken.
Pl: Cikkek kategorizálása, Kommentek elemzése, Bulling, Indirekt sértő kifejezések szűrése,
Fenyegető üzenetek szűrése ...

Köszönöm a figyelmet!