Adatbányászat prezentáció

Forman Balázs Attila és Michaletzky Tamás Vilmos

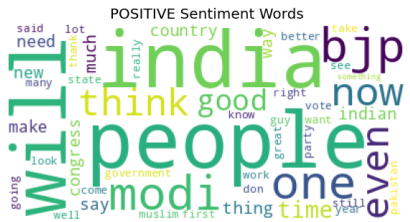
2021

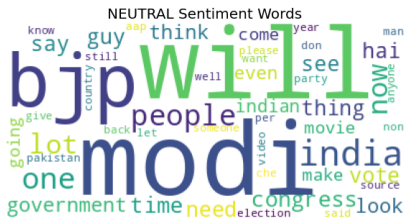
Adatbányászat

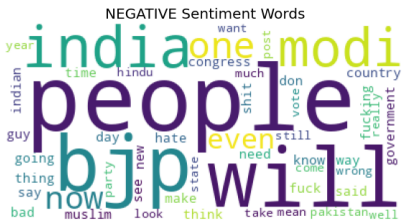
1. Bevezetés
   - Adathalmaz jellemzése
   - A probléma megfogalmazása, hipotézisek
   - Előfeldolgozás
2. Algoritmusok
   - TF-IDF
   - Word2Vec és Doc2Vec
   - GloVe
   - Bert
3. Összegzés
   - Összehasonlítás
   - Végszó

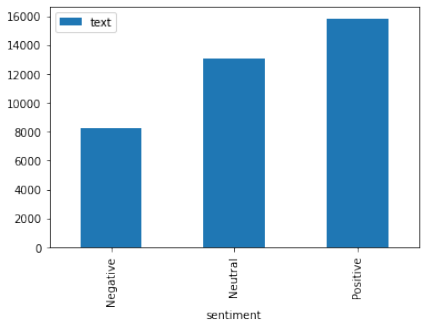
Adathalmaz jellemzése

- **Kaggle**-ről
- 37249 darab **Reddit** bejegyzés: HATE SPEECH
- -1, 0, és 1 számmal
  - negatív
  - semleges
  - pozitív









Hipotézisek

Feladat

Hate speech detection különböző NLP-módszerekkel

- Különböző hatékonyság

$$\texttt{TfIdf} < \texttt{word2vec} < \texttt{doc2vec} < \texttt{GloVe} < \texttt{Bert}.$$

- binary $>$ multiclass
- Pozitív eltolódás
- Dimenzió
  - TF-IDF nagy dimenzió
  - Többinek közepes
- Túltanulás félelme

Előfeldolgozás

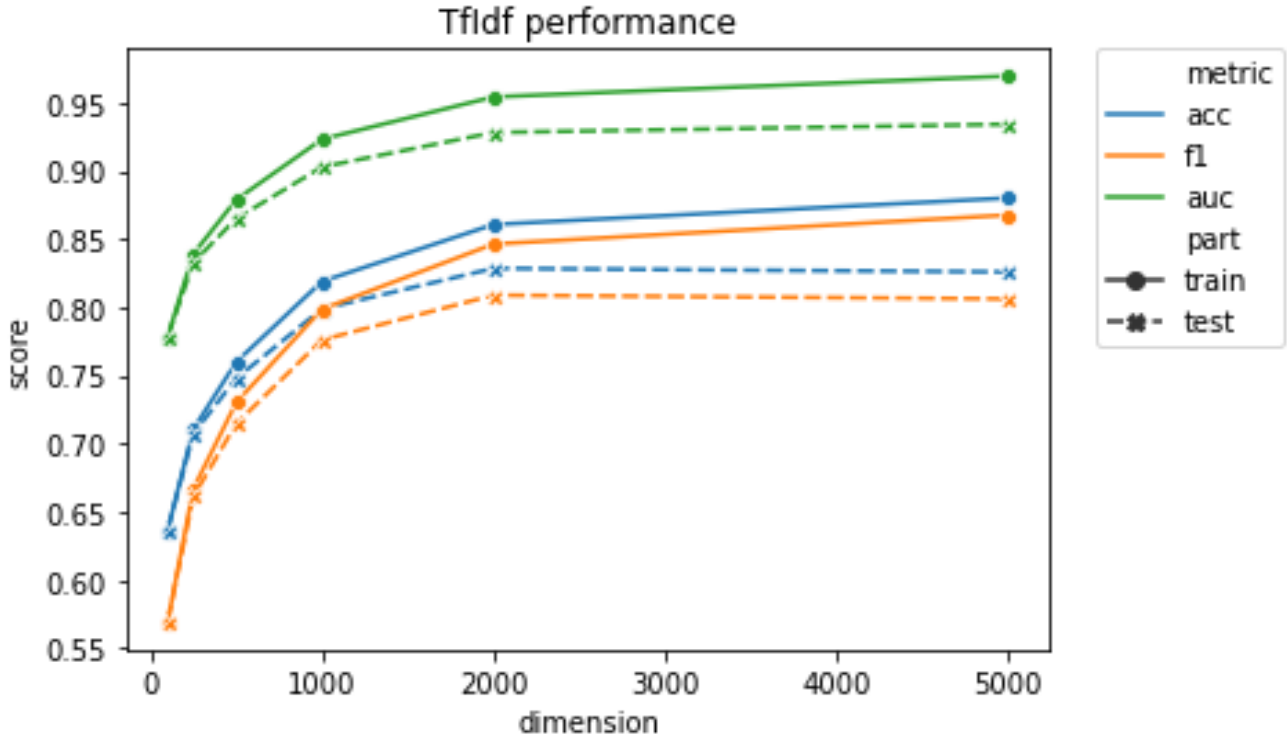
- Bejegyzések szavakra bontása
- Kisbetű
- Kötőszavak, nem latin betűk
- Szótövek: lemmatizálás, stemmelés

Példa

- Original tweet: family mormon have never tried explain them they still stare puzzled from time time like some kind strange creature nonetheless they have come admire for the patience calmness equanimity acceptance and compassion have developed all the things buddhism teaches
- Processed tweet: ['famili', 'mormon', 'never', 'tri', 'explain', 'still', 'stare', 'puzzl', 'time', 'time', 'like', 'kind', 'strang', 'creatur', 'nonetheless', 'come', 'admir', 'patienc', 'calm', 'equanim', 'accept', 'compass', 'develop', 'thing', 'buddhism', 'teach']

TF-IDF

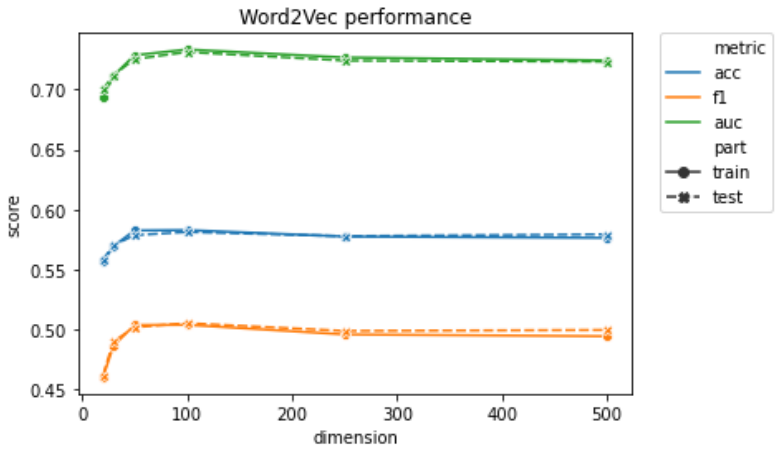
Hagyományos, szógyakoriság alapú módszer, a *term frequency, inverse document frequency* szavakból.

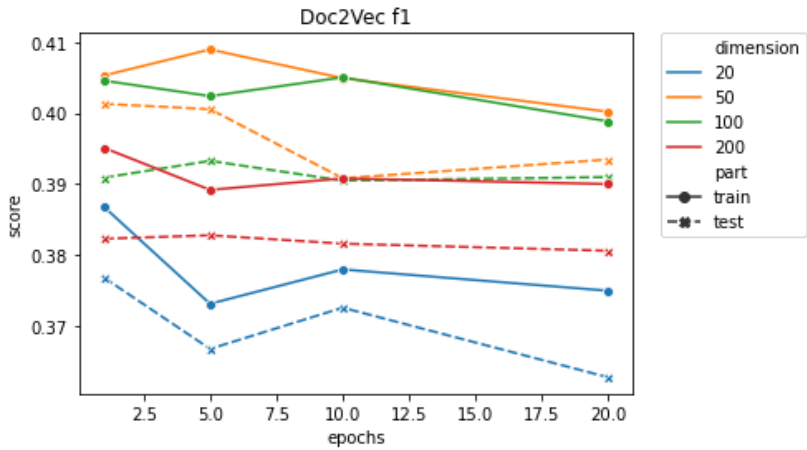


Word2Vec és Doc2Vec

Neurális háló alapú beágyazások.

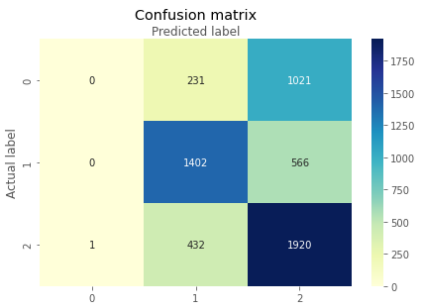
- word2vec: szavakat ágyaz be, hasonló szavak hasonlóan
- doc2vec: a dokumentumokat felelteti meg egy-egy vektornak, több epochon át javítva a beágyazáson

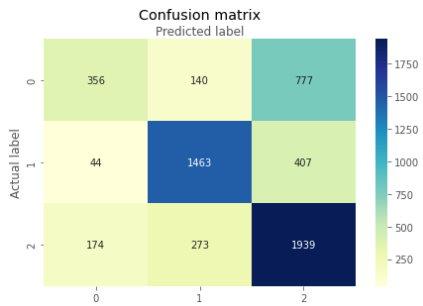


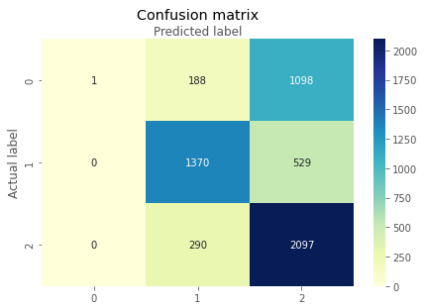


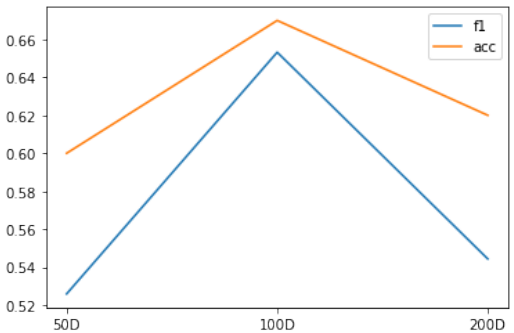
GloVe és LSTM

- A **GloVe** a **Stanford** egyetem 2014-es fejlesztése, a *Global Vectors for Word Representation* kifejezésből jön a neve.
- Felügyelet nélküli szóbeágyazó algoritmus
- Együttes előfordulási gyakoriságokra alapszik



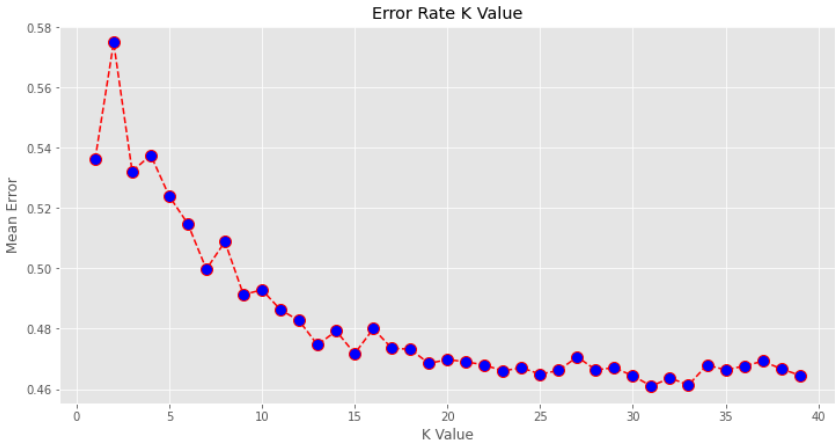


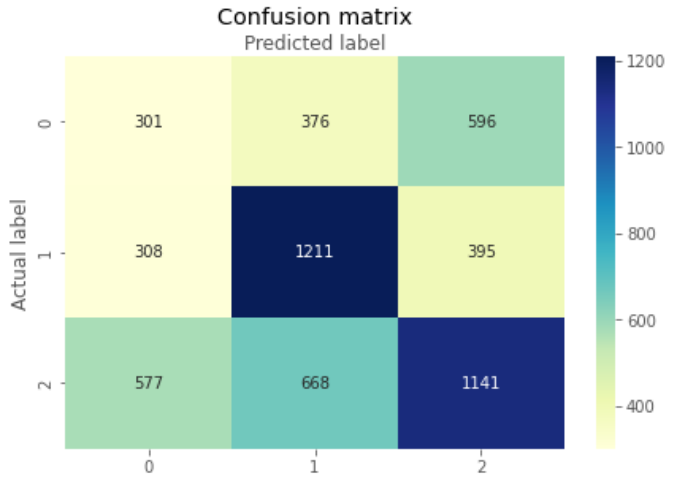




GloVe és K-neighbours

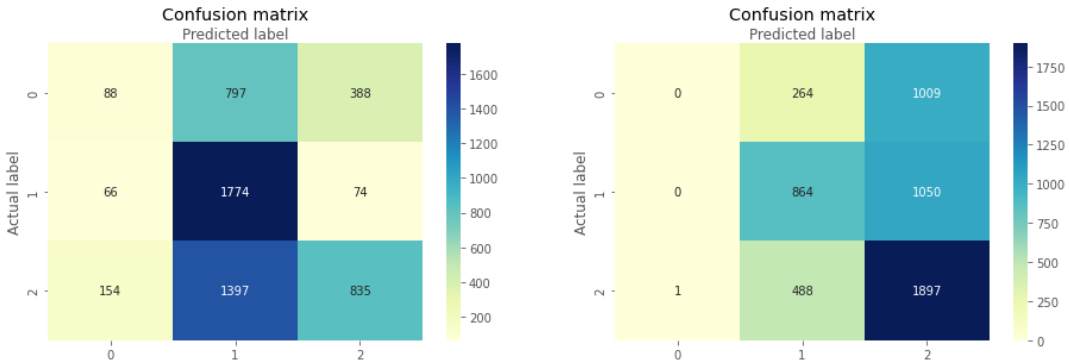
- 100 dimenzióba ágyaztunk be
- a több szomszéd nem vezetett lényegesen pontosabb jósláshoz
- nagyobb valószínűséggel mond 0-t

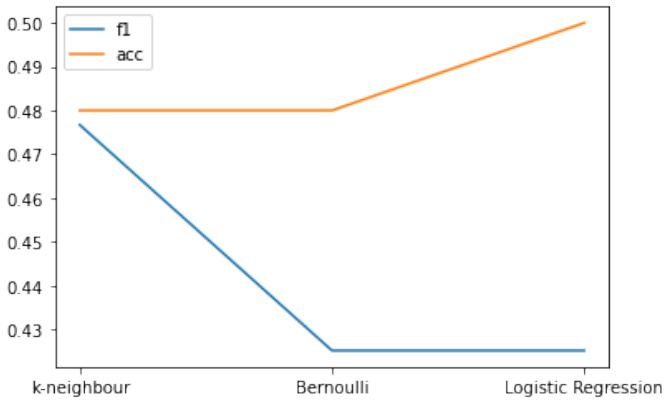




GloVe, Bernoulli és Logisztikus regresszió

- Ritkán jósolnak negatív szentimentet
- Hasonlóan pontosak, mint a *kNN*
- Itt is 100 dimenzióba ágyaztunk be





Bert

A Bert a Google 2018-as fejlesztése: *Bidirectional Encoder Representations from Transformers*

- sok NLP-feladat up-to-date legjobb megoldása: pl. szöveggenerálás

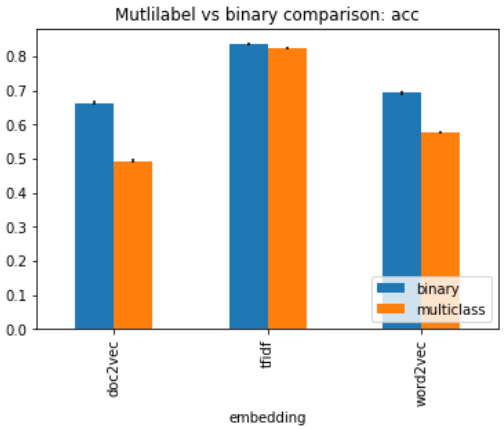
HuggingFace könyvtárból PyTorch alapon Bert-Medium: 8 réteg, 512 dimenzió

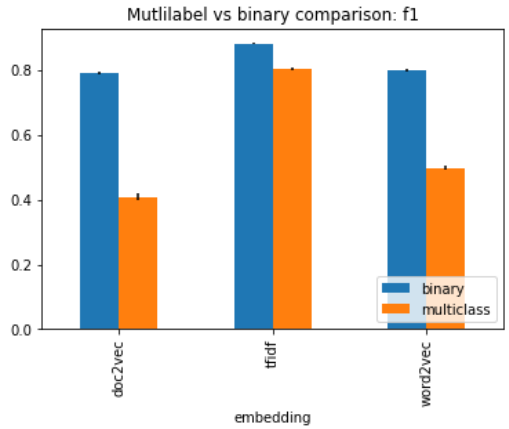
De a klasszifikálás is jól megy neki!

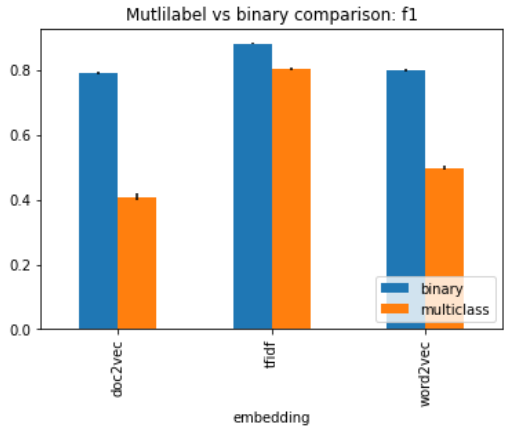
Összehasonlítás

Legjobb mérések és beállítások:

MULTICLASS	Dimenzió (epoch)	Accuracy	F1
Tfldf	2000	0.87	0.83
word2vec	50	0.58	0.50
doc2vec	50 (5)	0.48	0.40
GloVe + 5NN	100	0.48	0.47
GloVe + LSTM	100 (5)	0.67	0.65
Bert	512 (2)	0.92	0.92







Végszó

Hipotézisek

- különböző hatékonyság: MÁSKÉNT

$$\texttt{word2vec} = \texttt{doc2vec} < \texttt{GloVe} < \texttt{TfIdf} < \texttt{Bert}.$$

- binary $>$ multiclass: IGEN
- pozitív eltolódás: egyes módszerek IGEN, máshol NEM volt tapasztalható
- dimenzió
  - TF-IDF nagy dimenzió: 5000: IGEN
  - többinek közepes: 50: IGEN, sőt kicsi
- túltanulás félelme: NEM volt tapasztalható

Köszünjük a figyelmet!

- https://github.com/tmichaletzky/datamining2021