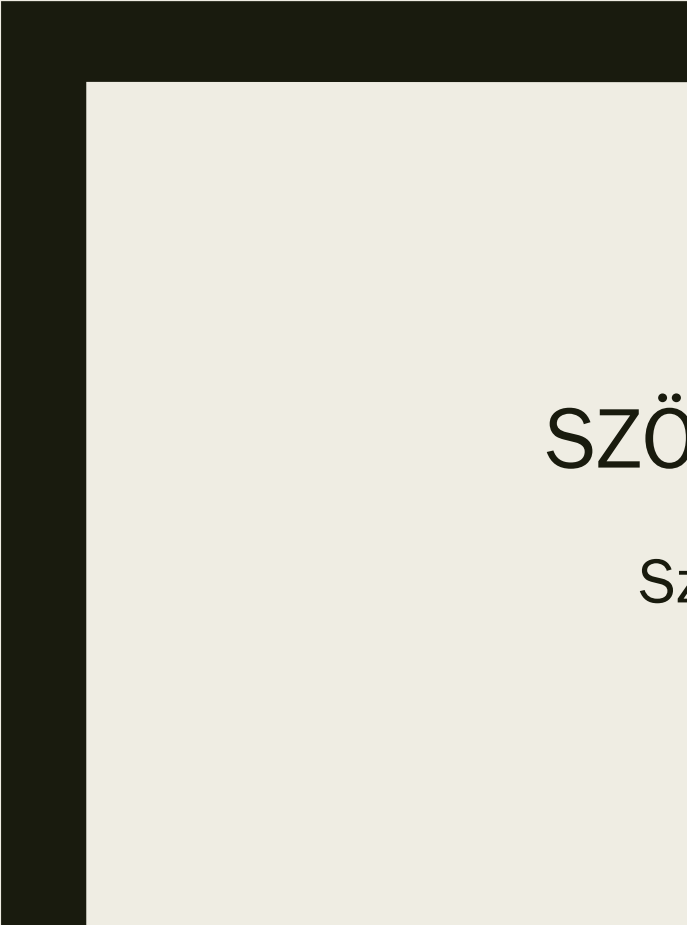




SZÖVEGANALITIKAI MODELLEK

Szerzők nem szerinti klasszifikálása

Bolonyai Flóra

f.bolonyai@gmail.com



kontextus

- Profilozás
 - *explicit vagy implicit*
 - **szöveges adatok alapján**
- Szerzőklasszifikáció
 - *azonosítás*
 - *látens tulajdonságok*

hogyan lehet írott szöveg alapján megállapítani a szerző nemét?

ELMÉLETI HÁTTÉR

- Miért van ilyesmire szükség?
- Miért gondoljuk, hogy ez lehetséges?
- Korábban milyen különbségeket találtak, akik ezt vizsgálták?
- Milyen lépésekből áll egy klasszifikációs modell illesztése?

SAJÁT MODELL

- Az adatbázis bemutatása
- A kiválasztott magyarázó változók és modellek bemutatása
- Az illesztés folyamata
- Eredmények

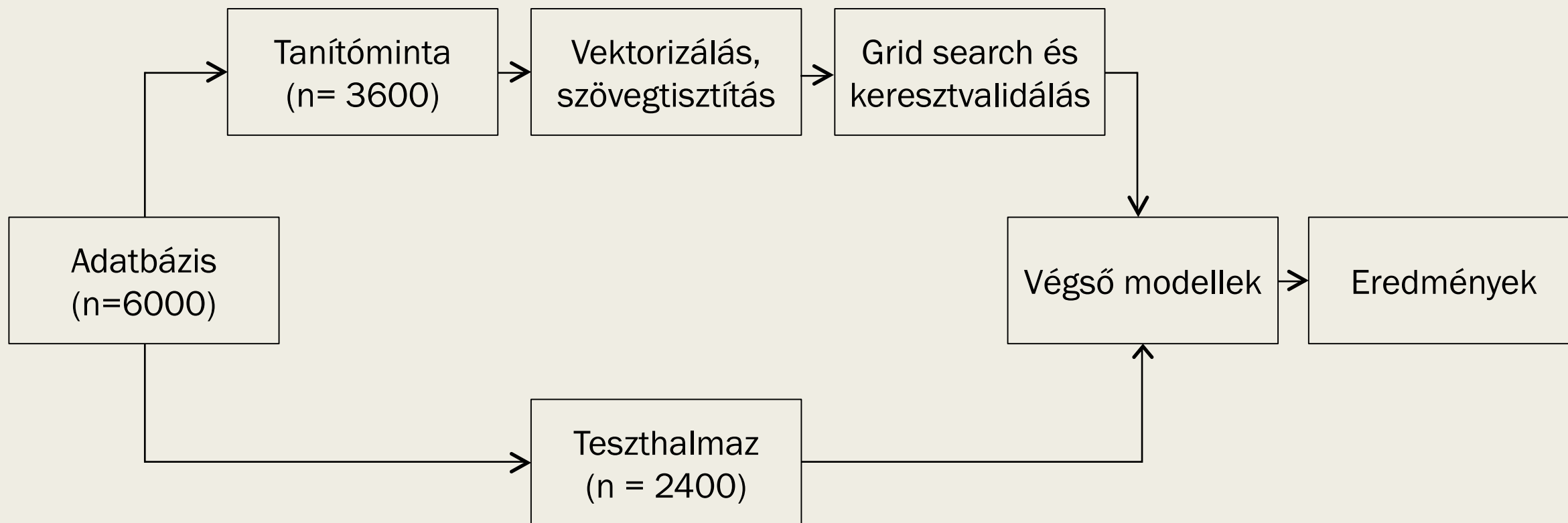
elméleti háttér I.

- a gender profiling felhasználási területei: marketing, bűnügyi nyelvészet, véleménybányászat
- Lakoff (1973): Language and Woman's Place
 - *empirikus bizonyítékok vagy cáfolatok keresése → vannak viszonylag egybevágó megállapítások*
 - *az elmélet megkérdőjelezése*
 - a nem inkább confounding változó
 - a nyelv nem bináris kategóriaként ad infót a nemről, hanem arra szolgál, hogy a beszélő pozícionálja magát a hallgatósághoz képest (Bamman et al.)
 - *nem lehet a nemmel magyarázni a különbségeket, de ettől még a klasszifikálás problémája izgalmas*

elméleti háttér II.

- klasszifikációs modellek a gyakorlatban
 - *korpuszok*
 - az utóbbi időkben dominálnak az internetről letöltött szövegek
 - *magyarázó változók (feature-ök)*
 - lexikális – szintaktikai – tartalmi – strukturális – karakteralapú
 - miről – hogyan
 - *modellek*
 - hasonlóság – egyszerű gépi tanulási modellek – neurális hálók

saját elemzés



eredmények

- SVM és logisztikus regresszió: 82%
- a lineáris modellek jobban működnek, mint a random forest
- a legpontosabb modellekben nagyon sok a feature
- más korpuszokon sem reménytelen
- legfontosabb változók meghatározása



hibás jóslatok

- mennyiben hasonlítanak az egyes modellek által rosszul besorolt szövegek?
- kvalitatív vizsgálat → domináns témák

férfi felhasználó, LR becslés: $P(\text{férfi})=0,02$, RF becslés: $P(\text{férfi})=0,32$

<I have been fighting with ingrown hair on my neck for over a month 😞 it's still there 😡 Any advise how to get rid of it?>

<Okay so I'm not fitting in size 30 shorts anymore 😞>

<I want a tattoo and I have an amazing idea. I might need bigger legs 🙄>

férfi felhasználó, LR becslés: $P(\text{férfi})=0,01$, RF becslés: $P(\text{férfi})=0,38$

<When your new job turns out to be just as shit as the one you left 😞😞😞>

<Been waiting all morning for the carpet driers and glassie to turn up to fix earthquake damage
Going to get a coffee. Bet they both turn up>

<My friend tagged me in a post about a Harry Potter wand remote for TVs. Cracked her up when I admitted I already own it 😞😞>

női felhasználó, LR becslés: $P(\text{férfi})=0,99$, RF becslés: $P(\text{férfi})=0,73$

<@user Klopp's talked bout importance of cups, team selection could suggest priorities elsewhere - back in CL, better calibre recruitment>

<Kinda glad for @user that that match ended 1-0 cos it means his first goal for 7 years was the winner! #PLYLIV>

<Wonder what odds you'd have got on Everton inflicting Pep Guardiola's biggest ever league defeat? #EVEHCI>

női felhasználó, LR becslés: $P(\text{férfi})=0,92$, RF becslés: $P(\text{férfi})=0,68$

<So here's the question: do sanctions no longer apply when a Russian entity is taking an action to comply with a request from the FSB?>

<#FutureHistoryBooks: "By the middle of the 2nd week, Flynn decided he had waited long enough, and took steps to speed up his war with Iran.">

<Trump's own campaign site still announces that he is "calling for a total and complete shutdown of Muslims entering... url">

folytatás?

- lehetséges-e a pontosabb klasszifikálás?
- más látens tulajdonságok megállapítása
- más nyelvek bevonása
- a nyelvhasználatot befolyásoló egyéb mutatók bevonása