Lietuvos mokslo taryba

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Seminaras - diskusija „Socialinės skirtys Lietuvoje: ką rodo Tarptautinės socialinio tyrimo programos duomenys? “

Automatinė interneto komentarų sentimentų analizė

Jurgita Kapočiūtė-Dzikienė

Projektą „*Tarptautinė socialinio tyrimo programa: Lietuvos socialinių problemų stebėsena” (ISSP-LT)* “ finansuoja Lietuvos mokslo taryba (sutarties Nr. SIN-07/2012)

Sentimentų analizė

- Sentimentų analizė padeda nustatyti kalbėtojo
  - teigiamą,
  - neigiamą,
  - neutralų

  požiūrį aptariama tema

- Sentimentų analizės uždavinys sprendžiamas lietuviškiems interneto komentarams

Aktualumas

- Bendriems uždaviniams:
  - kompanijos sužino nuomonę apie jų prekes/paslaugas
  - sociologai sužino požiūrį į visuomenei reikšmingus įvykius
  - psichologai analizuoja emocijas ir jų atsiradimo priežastis
  - ir t.t.

- Specifiniams uždaviniams:
  - grasinimo turinio žinučių aptikimui
  - įžeidžiančio turinio žinučių iš konkretaus asmens išsiuntimo prognozavimui
  - ir t.t.

Susiję darbai: kalbos faktorius

- Daugybė įvairių metodų (daugiausiai anglų kalbai)
- Metodo pasirinkimą įtakoja kalbos savybės
- Sentimentų analizės uždavinys lietuvių kalbai anksčiau nebuvo spręstas

Lietuvių kalba

- Stipriai kaitoma
- Turtinga kalbdaros sistema:
  - pvz.: 78 priesagos mažybiniams/maloniniams žodžiams; 25 daiktav., 19 veiksmaž. priešdėlių ir t.t.
- Laisva sakinio struktūra:
  - pvz. frazė tu esi labai geras turi didesnį teigiamą atspalvį, nes akcentuojamas žodis labai; lyginant su tu esi geras labai kurioje akcentuojamas geras
- Neiginio traktavimas:
  - Priešdėlis ne- arba nebe- keičia žodžio prasmę: (ne)geras, (ne)dirbti, (ne)laimė
  - Ne, nebe, nėra einančios atskirai keičia prasmę – t.y. išreiškia prieštaravimą: ne blogas, o geras
  - Negatyvi mintis išreiškiama keliais žodžiais niekas gerai nežaidžia (anglų k. vienu žodžiu: nobody plays well)

Duomenų aibė

- Interneto komentarai:
  - po “Lietuvos ryto” (www.lrytas.lt) straipsniais (iš aktualijų skilties)
- Duomenų aibės specifika:
  - neformali kalba: žargonas, barbarizmai, užsienio kalbų intarpai
  - diakritininių ženklų nepaisymas: ą, č, ę, ė, į, š, ų, ū, ž → a, c, e, e, i, s, u, u, z.

Kuo sudėtingas duomenų aibės anotavimas?

- Komentarai ištraukti automatiškai
- Anotuoti rankiniu būdu, bendru 2 anotuotojų sutarimu
- Sudėtinga nes:
  - ankstesni interneto komentarai įtakoja esamo teksto prasmę /neatsižvelgta/
  - susipynusios įvairios nuomonės /tokie komentarai šalinti/
  - sarkazmas, kurį kartais sunku atpažinti /anotuota/

Duomenų aibės statistika

komentarai	komentarų kiekis	žodžių kiekis	skirtingų žodžių kiekis
teigiami	1.500	10.455	6.394
neigiami	1.500	15.000	7.827
neutralūs	1.500	13.165	4.039
viso	4.500	38.621	15.008

- Subalansuota duomenų aibė
- Atsitiktinis tikslumas ir didžiausios klasės tikimybė: 0,333

Susiję darbai: žodyninis metodas

- Būdvardžiai – patys populiariausi sentimentų indikatoriai

- Būdvardžiai + prieveiksmiai duoda daug geresnius rezultatus nei vien tik būdvardžiai (Benamara ir kt., 2007)

- Būdvardžiai + prieveiksmiai + daiktavardžiai + veiksmažodžiai svarbūs sentimentų analizei (Taboada ir kt., 2011)

Darbe naudotas žodyninis metodas

- Žodyno žodžiai susieti su poliškumą nusakančia reikšme:
  - **-3** – stipriai neigiamas
  - **-2** – vidutiniškai neigiamas
  - **-1** – silpnai neigiamas
  - **0** – neutralus
  - …
  - **+3** – stipriai teigiamas

- Tekste atpažįstami žodžiai, nustatomos jų reikšmės ir apskaičiuojamas bendras rezultatas:
  - **>0** – komentaras teigiamas
  - **<0** – neigimas
  - **=0** – neutralus

Kuo sudėtingas žodyno rengimas?

- Būdvardžiai, prieveiksmiai, daiktavardžiai ir veiksmažodžiai ištraukti automatiškai (iš ~1 mln. žodžių lietuvių kalbos tekstyno)

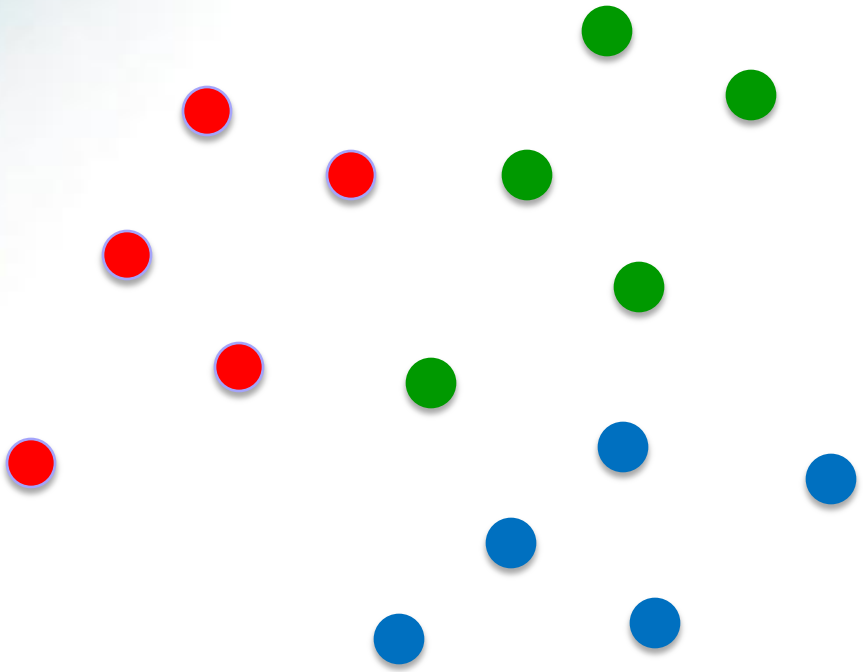
- Anotuoti rankiniu būdu, bendru 2 anotuotojų sutarimu

- Žodynas automatiškai papildytas sinonimais iš Lietuviško WordNet

Sukurto žodyno statistika

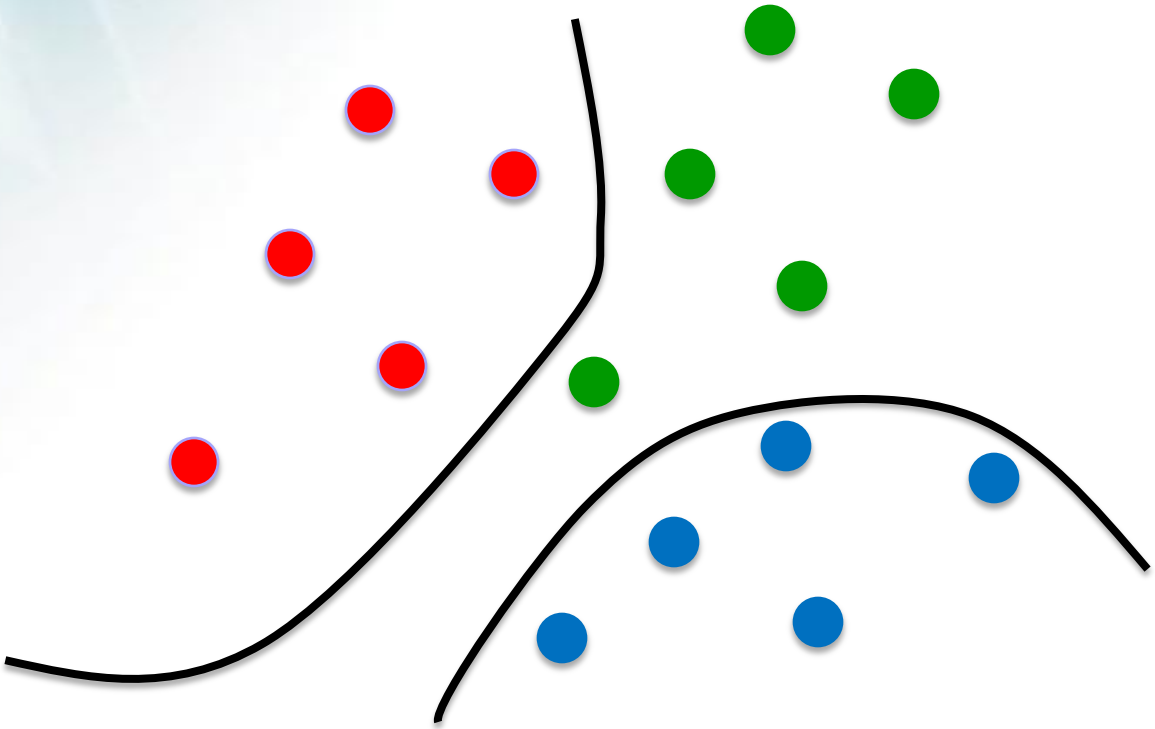
Poliškumo reikšmė	Būdv.	Priev.	Veiksmaž.	Daiktav.	Viso
-3	138	74	236	296	744
-2	175	122	337	775	1,409
-1	267	95	733	1.945	3.040
0	4.392	1.362	10.039	12.719	28.512
1	163	122	344	896	1.525
2	148	117	113	213	591
3	142	62	72	55	331
Viso	5.425	1.954	11.874	16.899	

Mašininio mokymo metodo principas



- Komentarai: teigiami, neigiami, neutralūs.

Mašininio mokymo metodo principas



- Komentarai: teigiami, neigiami, neutralūs.

Susiję darbai: teksto savybės MM metodams

- Mašininiame mokyme (MM) dažniausiai atsižvelgiama į pavienius žodžius (Pang ir kt., 2002)
- Žodžių junginiai (iki 3 žodžių) veikia tiksliau nei pavieniai žodžiai (Dave ir kt., 2003); junginiai po 3, 4, 5, 6 žodžius veikia tiksliau nei pavieniai žodžiai ar žodžių 2-etai (Cui ir kt., 2006); žodžių 2-etai veikia tiksliau nei pavieniai žodžiai ar žodžių 3-etai (Pak ir Parubek, 2011)
- Žodžių kamienai veikia tiksliau nei pavieniai žodžiai (Dave ir kt., 2003)
- Simbolių grandinėlės veikia tiksliau negu žodžių junginiai (Hartmann ir kt., 2011)

Naudotos teksto savybės su MM metodais

- Žodžiai (dažniausiai naudojama)
- Žodžiai + žodžių 2-jetai (aukštesnės eilės junginiai kartais veikia tiksliau nei pavieniai žodžiai)
- Lemos (rekomenduojama stipriai kaitomoms kalboms)
- Simbolių po 4 grandinės, pvz.: žemės ūkis → žemė, emės, mės_, ės_ū, s_ūk, _ūki, ūkis (geriausias klasifikavimui į temas (Kapočiūtė-Dzikienė ir kt.))

Pirminiai teksto apdorojimo būdai

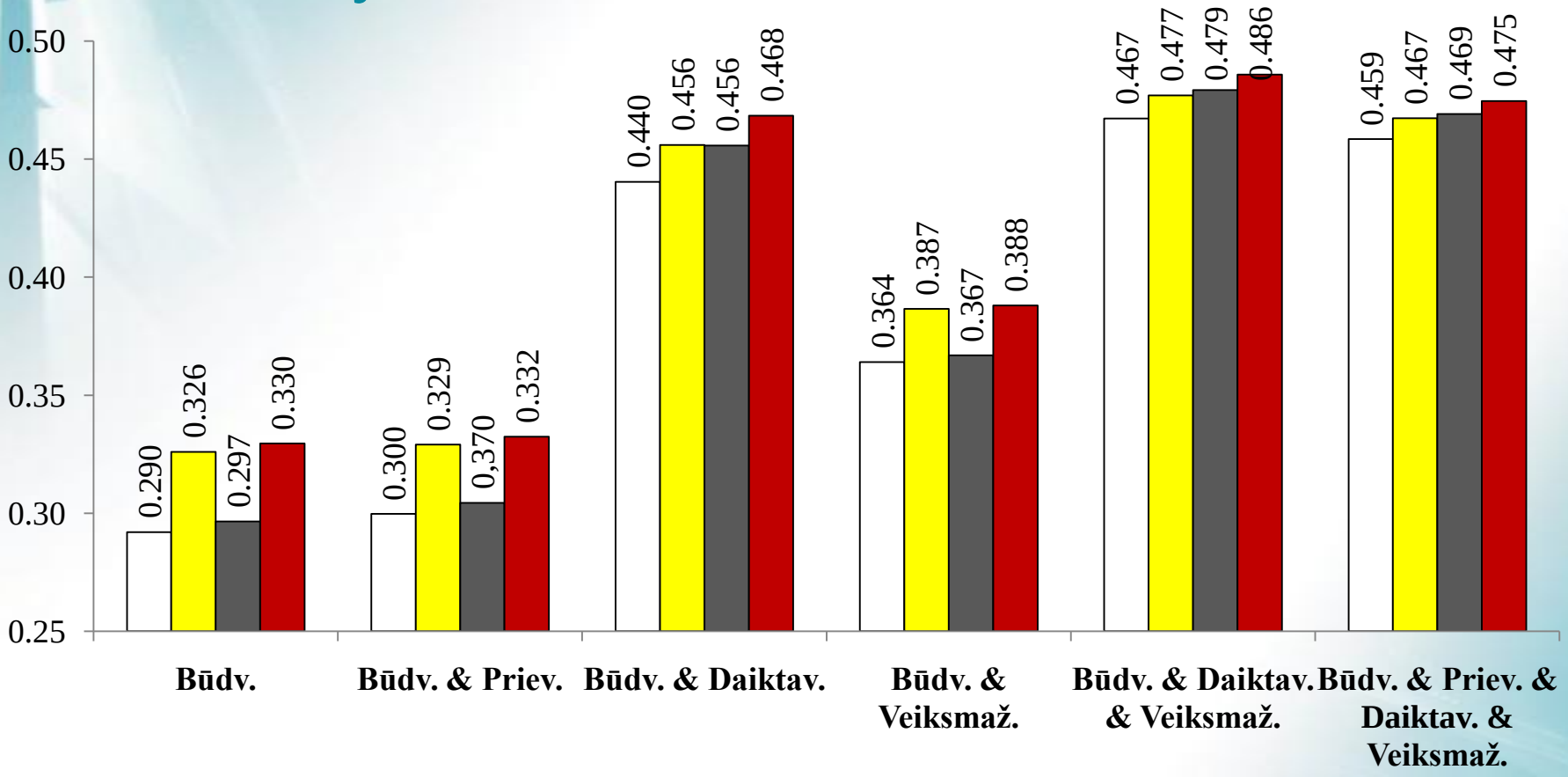
- lemavimas (naudojant Daudaravičius ir kt. (2007) kurtą lemavimo įrankį lietuvių kalbai)
- jausmus nusakančių simbolių keitimas jausminiai žodžiais (sudarytas 32 jausminių žodžių sąrašas
- diakritinių ženklų šalinimas (ą, č, ę, ė, į, š, ų, ū, ž keičiami atitinkamais a, c, e, e, i, s, u, u, z)
- trumpų žodelių šalinimas (išskyrus jaustukus (Spencer ir Uchyigit, 2012))
- Raidžių keitimas mažosiomis, skaičių ir skyrybos ženklų šalinimas

Pirminių teksto apdorojimo būdų įtaka*

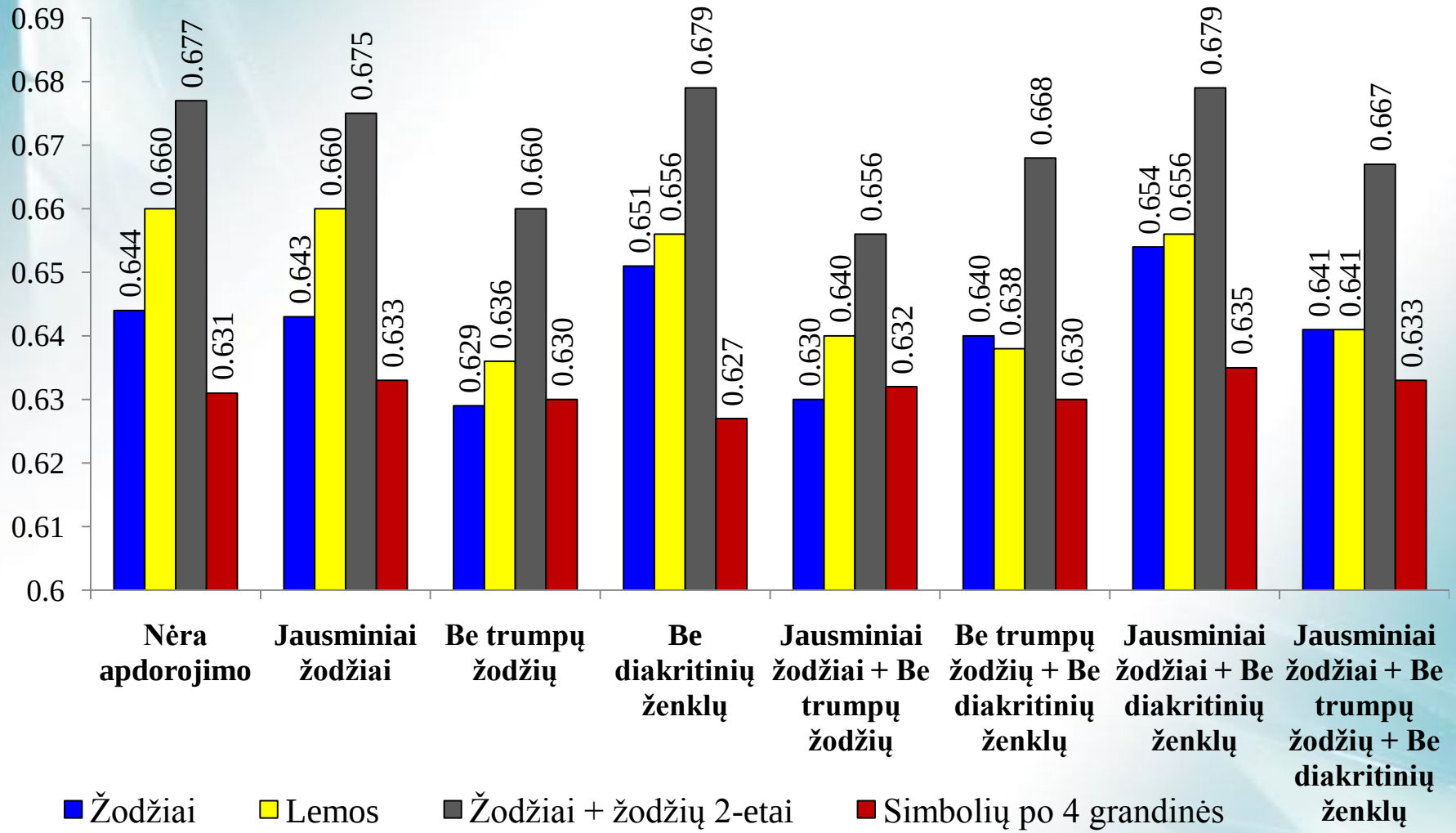
komentaras	Nėra apdorojimo	Po lemavimo	Po jausminių žodžių terpimo	Po trumpų žodelių šalinimo	Po diakritinių ženklų šalinimo
teigiami	10.455 6.394	10.386 3.177	10.664 4.027	8.982 3.941	10.455 3.724
neigiami	15.000 7.827	14.928 6.475	15.107 7.811	11.945 7.716	15.000 7.457
neutralūs	13.165 4.039	13.084 5.134	13.226 6.391	10.427 6.276	13.165 6.058
viso	38.621 15.008	38.398 11.669	38.997 14.966	31.354 14.923	38.621 13.983

* Pirma reikšmė nusako visų žodžių kiekį, antra – tik skirtingų

Žodyninio metodo tikslumas



MM metodo tikslumas



Klaidų analizė

- Prieveiksmiai nėra geri sentimentų indikatoriai lietuvių kalbai:
  - dažnai neišreiškia jokių sentimentų, pvz.: gerai pasakyta
  - negatyvūs žodžiai sustiprina pozityvią emociją, pvz.: žiauriai geras
- Žodyninis metodas nėra efektyvus sentimentų identifikavimui
- Diakritiniai ženklai vis dar yra problema: pvz. sauni nebus atpažinta ir sulemuota kaip šaunus
- Sentimentai gali būti išreiškiami netiesiogiai, pvz. Auksas! Valio!
- Sarkazmas, pvz. Peilis puikus, tik po antro naudojimo sulūžo.

Išvados

- Diakritinių ženklų šalinimas ir jausminių žodžių terpimas – geriausios teksto apdorojimo technikos
- Būdvardžiai, daiktavardžiai ir veiksmažodžiai naudojami drauge yra geriausi sentimentų indikatoriai taikant žodyninį metodą
- Mašininio mokymo klasifikavimo metodas veikia geriau nei žodyninis
- Geriausia teksto savybė (naudojama su mašininio mokymo metodu) – žodžiai ir žodžių 2-ejetai

- **Didžiausias pasiektas interneto komentarų klasifikavimo tikslumas lietuvių kalbai yra ~0.679**

Klausimas sociologams

- Ar toks tikslumas jums pakankamai geras, kad galėtumėte taikyti savo tyrimams?