Cybernetica AS

Infoturbeinstituut

Andmeteaduse ja tekstikaevemeetodite rakendamine eestikeelsete meediaväljaannete näitel

Praktikaaruanne

Praktikant: Anne-Liis Rämson

Juhendaja: Jan Willemson, PhD

Tartu

2021

Lühikokkuvõte

Praktikaprojekti eesmärgiks on uurida kolme eestikeelse meediaväljaande (Delfi, Telegram, Uued Uudised) põhjal uudisartiklite tekste programmeerimiskeele Python abil. Esmalt koguti andmed veebiämblike ehk *crawler*'ite abil. Uudisartiklite tekstid kirjeldati sõnapilvede ja sagedasemate sõnade loendite abil, kasutades sõnaliikide määramiseks teegi EstNLTK vahendeid. Uudisartiklite iseloomustamiseks kasutati meelsusanalüüsi, toetudes Eesti Keele Instituudi koostatud emotsioonileksikonile, leiti iga artikli kohta artikli negatiivsust näitav kordaja. TF-IDF meetodit kasutades jagati uudisartiklid tekstide põhjal kolme klastrisse. Projekti viimases osas rakendati nelja juhendatud masinõppemeetodit uudisartiklite valedetektori loomiseks. Uudisartiklite tekstide sõnestamiseks kasutati eesti keele sõnestamise vahendeid.

Application of data science and text mining methods in the example of media publications in Estonian

Abstract

The aim of the internship programme is to examine news texts of three Estonian media publications (Delfi, Telegram, Uued Uudised) using the programming language Python. In the first, data was collected using web crawlers. News texts were described using word clouds and the most common words lists using EstNLTK library tools to determine part of speech. Sentiment analysis was used to characterize the news articles with emotion lexicon of the Estonian language formed by EKI and to detect the coefficient indicating the negativity of the news article. Using TF-IDF method news articles were divided into three clusters based on news texts. In the last part of the project four supervised machine learning methods were implemented to create a fake news detector for news articles. Estonian word tokenizing tools were used to tokenize news texts.

Sisukord

1. Sissejuhatus

Praktika toimus Cybernetica AS infoturbeinstituudi Tartu kontoris ajavahemikul 07.09.2020 – 31.12.2020. Praktika sooritati Tartu Ülikooli infotehnoloogia mitteinformaatikutele magistriõppe kursuse Praktika informaatikas (MTAT.03.206, 12 EAP, sessioonõpe, kestvus 312 töötundi) raames. Praktika juhendaja oli Jan Willemson.

Praktika peamised eesmärgid

- saada ülevaade väärinfo levitamisest kui ühest infoturbe probleemist;
- arendada algoritmilist mõtlemist ja süvendada programmeerimise oskusi;
- katsetada väärinfo ja tõese info põhjal tekstianalüüsi kasutades EstNLTK teegi võimalusi eestikeelsete tekstide töötlemiseks;
- luua detektor väärinfo tuvastamiseks;

Projektiga seotud failid asuvad: https://github.com/anneliisramson/eestikeelsed-uudised.

2. Mõisted

Eestikeelsetes tõlkeartiklites on terminit *„fake news“* tõlgitud nii võltsuudisteks, libauudisteks kui ka valeuudisteks (Klaassen, 2018).

“**Desinformatsioon** ehk tõendatavalt väär või eksitav teave, mida luuakse, avaldatakse ja levitatakse majandusliku kasu eesmärgil või üldsuse ettekavatsetud petmiseks, moonutab avalikku arutelu, õõnestab kodanike usaldust institutsioonide ja meedia vastu ning isegi destabiliseerib demokraatlikku protsessi (näiteks valimisi).” (Euroopa Liidu Infokeskus)

Valeuudis tähistab uudist, mis sisaldab võltsitud ja/või eksitavat materjali, kuid mida auditoorium peab tõeseks, kontrollitud faktidega usaldusväärseks uudiseks (Himma-Kadakas, 2017).

Libauudis

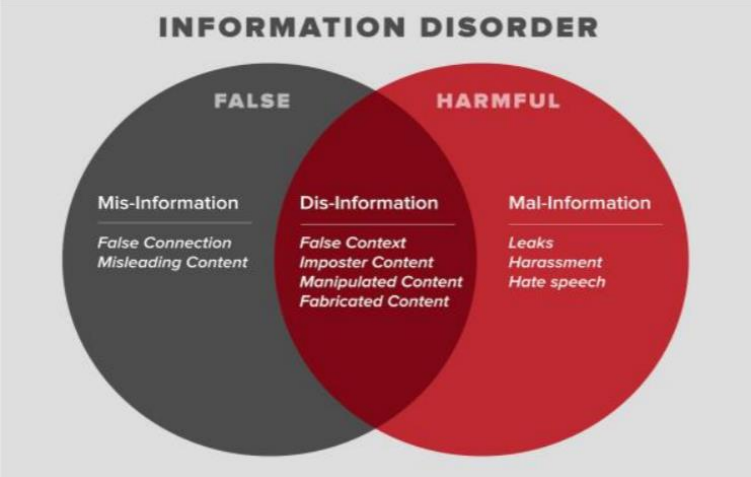
Libauudise sisu peaks auditooriumile teadaolevalt mitte tõetruu olema, näiteks 1. aprilli uudised või rubriigid Ekspressi Kranaat, Rohke Debelaki uudised jt.

Võltsuudis

Võltsuudist eristab valeuudisest tahtlikkus – valeuudis võib sisaldada kogemata valeinformatsiooni, võltsuudis on sihilikult petmiseks või eksitamiseks loodud uudis (Klaassen, 2018).

Infokorratus

- eksitav info (*Mis-Information*) – valeinfo levitamise eesmärk ei ole sellega kahju tekitada;
- valeinfo (*Dis-Information*) – infot levitatakse teadlikult ja tahtlikult kahju tekitamise eesmärgil ;
- vaenulik info (*Mal-Information*) – põhineb tegelikkusel, info kasutamise eesmärk on kahjustada persooni, organisatsiooni või riiki **(**Wardle & Derakhshan, 2017).



Joonis 1. Infokorratus **(**Wardle & Derakhshan, 2017).

Kõvad ja pehmed uudised

Hennoste on raamatus (Hennoste, 2008) esitanud *International Press Institute´*i uudiste jaotuse:

- kõvad uudised – sõda, poliitika, välissuhted, riigikaitse, majandus;

- pehmed uudised – kultuur, haridus, teadus ja tehnika, sotsiaalsed sündmused, *human interest* (üldhuvi teemad), kuritegevus, õnnetused ja kuritööd, sport, usk jne.

Kõvad uudised esitavad informatsiooni maailmas asetleidvate sündmuste kohta ja aitavad inimestel maailma paremini mõista. Nende teemad on tähtsad paljudele inimestele, nad on seotud avaliku eluga, milles toimuvad sündmused mõjutavad paljude inimeste elu ning nende oluliseks omaduseks on värskus ja objektiivsus.

Pehmeid uudiseid on huvitav lugeda, kuid nad ei pruugi olla tähtsad ja inimeste elu mõjutavad. Pehmed uudised on tihti seotud eraeluga, tuntud inimestega, tihti pole pehmed uudiseid kuigi värsked. Selliste uudiste ülesanne on mõjutada emotsioone ja pakkuda inimesele meelelahutust (Hennoste, 2008).

Praktikaaruandes kasutatakse termineid valeuudis ja väärinfo tähistamaks meelega ja/või kogemata antud valeinfot. Uudisartiklitest on vaatluse all kõvad uudised – sise- ja välispoliitika, majandus, vaadeldakse ka teadusuudiseid ja arvamuslugusid.

3. Tõejärgne ühiskond

Tõejärgsel ehk tõepõhjata ajastul on üha raskem vahet teha tõel ja valel, (sotsiaal)meedias on info üleküllus, on tekkinud inforuumid, millel on kindel tarbijaskond. Taolistes inforuumides levib ja võimendub info, mille tõesust tagavad sama inforuumi tarbijad. Inforuumiväliste ekspertide väidetele tähelepanu ei pöörata või kuulutatakse ekspertide väited kiiresti ebatõeks.

Tõejärgset ühiskonda defineeritakse kui ajastut, kus uudiste tarbijad usuvad suurema tõenäosusega infot, mis ühtib nende isiklike arvamustega ning ei otsi ega vaja enam infot, mille õigsust on kinnitanud mitu erinevat sõltumatut allikat (Cooke, 2017).

Näidetena sellistest inforuumidest võib tuua lameda maa teooria pooldajad, koroonaviiruse eitajad, maskikandmise eitajad, MMS-i pooldajad, vaktsiinivastased.

Ekspress Meedia uuriva toimetuse juures tegutsev Eesti Päevalehe Faktikontroll jälgib Eesti avalikus inforuumis avaldatud väidete täpsust. Faktikontrolli tehes jälgitakse väite faktipõhisust, faktikontrolli läbipaistvust ja erapooletust (Ekspress Meedia Faktikontroll).

Eesti Päevalehes avaldas Faktikontroll 31.12.2020 artikli „TOP 5 | Milliseid valesid jäid Eesti inimesed tänavu uskuma? Maskivalede ning ebaravi lubaduste õnge langeti ohtralt„.

Artiklis nimetati TOP5 valedeks Eesti meedias 2020. aastal:

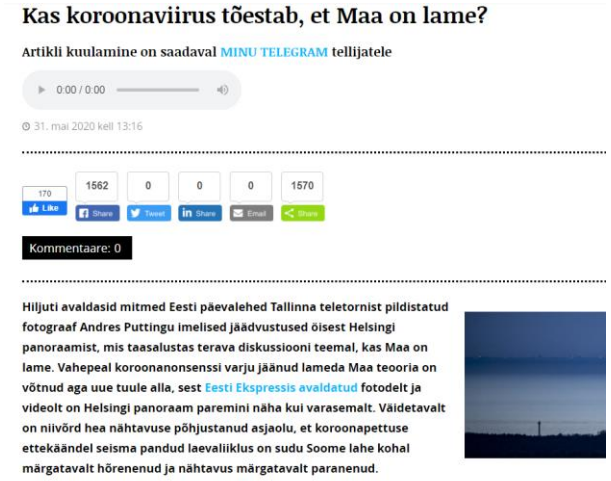
1. Tallinn läheb lukku (kevadise eriolukorra ajal).
2. Koroonapandeemia eitamine.
3. Maski kandmise eitamine.
4. Pseudokirjanduse võidukäik (nt Telegrami „Koroona valehäire“).
5. Eestiga seonduvad uudised Venemaalt (EPL, 31.12.2020).

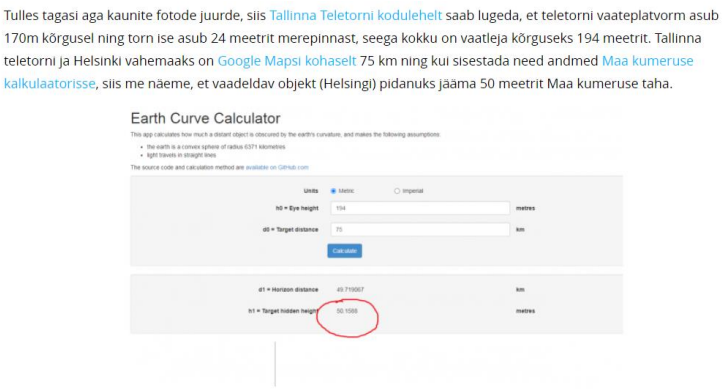
4. Kahe uudise võrdlus. Kas maa on lame?

Järgnevalt on kõrvutatud kaht sarnast uudist. Eesti Ekspressis ilmus 29.04.2020 uudis üle Soome lahe nähtavuse paranemisest. Fotodega tõestatakse, et Tallinnast on hästi näha Soome rannik ja vastupidi, sudu vähenemise põhjuseks tuuakse reisilaevade liikluse hõrenemine koroonakriisi tõttu.

Joonis 2. Lõik uudisest, Eesti Ekspress, 29.04.2020.

Telegramis ilmus 31.05.2020 samasisuline uudis, kuid hea nähtavuse põhjuseks tuuakse maa lamedus, lugeja saab tutvuda arvutustega, mis veenavad maa lapikuses.





Joonis 3. Lõik uudisest, Telegram, 31.05.2020.

Mõlemad uudised näivad usaldusväärsed ja tõesed, info tõesus tagatakse usaldusväärse fotograafi fotodega, sudu vähenemine Soome lahe kohal on igati usutav esimese koroonalaine ajal. Telegram täiendab lisaks eelpooltoodule uudist Maa kumeruse kalkulaatoriga, jättes õhku küsimuse, kas on võimalik näha objekte, mida ei tohiks näha olla. Kuid keda uskuda, kes räägib tõtt, kes valetab?

5. Projekti eesmärgid

Otsida uudistekstide tekstipõhist analüüsi tehes seaduspärasusi, sarnasusi ja erinevusi mitmete väljaannete lõikes. Leida seaduspärasusi, mis iseloomustaks tõest infot ja väärinfot. Leida vastust küsimusele, kas on võimalik uudisartiklitele tekstipõhist analüüsi rakendades eristada tõest infot ja väärinfot.

Eesmärkide täitmiseks tehti järgmist:

- koguti uudisartiklite andmestik;
- analüüsiti eestikeelsete meediaväljaannete tekste EstNLTK vahendeid kasutades;
- leiti seaduspärasusi erinevate väljaannete uudiste tekstidest;
- leiti võimalusi vale ja tõe eristamiseks tekstipõhiselt;
- töötati välja erinevaid meetodeid tekstide analüüsiks.

6. Uudisartiklite linkide kogumine veebist

Uudisartikleid koguti kolmest väljaandest:

- Delfi arhiiv;
- Telegram arhiiv;
- Uued Uudised arhiiv.

Valiku tegemisel oli määrav väljaannete erinäolisus ja teemade kattuvus. Uuritavate väljaannete arvu on võimalik edaspidistes uuringutes suurendada.

Linkide kogumiseks kasutati Pythonis kirjutatud Scrapy raamistikku veebist andmete „kraapimiseks“. Veebist uudislinkide kogumiseks vajalikud failid asuvad projekti *GitHub*'i keskkonna kaustades: *delfiscraper*, *telegram*, *uued_uudised_scraper*.

Järgnevas tabelis on väljaannete kaupa uudiste linkide valiku kriteeriumid. Väljaanded Telegram ja Uued Uudised ei võimalda valida artikleid ajaperioodi järgi. Rubriikide valimisel on jälgitud sarnaste teemade hõlmamist.

Väljaanne	Linkide valiku kriteeriumid	Linkide arv
Delfi arhiiv	periood: 01.01.2020 – 03.12.2020 kanal: Delfi kategooria: päevauudised	7497
Telegram arhiiv	Rubriigid: • Arvamus • Eesti • Maailm • New World Order • Teadus- ja tulevik	6439
Uued Uudised arhiiv	Rubriigid: • Arvamus • Eesti • Maailm • Majandus	14 417
Kokku linke		28 353

Tabel 1. Uudiste linkide valiku kriteeriumid.

Kogutud lingid salvestatakse csv-failidena.

7. Uudiste sisu kogumine – andmekorje

Kolme meediaväljaande lehtedelt kogutakse sarnasel viisil andmed artiklite kohta. Mõningad erisused andmete kogumisel tekivad veebilehtede erineva ülesehituse tõttu.

Andmete kogumine toimub *Jupyter Notebook* failides:

- *02 delfi_andmetekorje.ipynb;*
- *02 telegram_andmekorje_rubriigid.ipynb;*
- *02 uueduudised_andmekorje_rubriigid.ipynb.*

Enne andmete kogumist eemaldatakse Delfi puhul selle projekti jaoks mittevajalikud lingid (krimiuudised, kultuuriuudised, „Kroonika“ uudised, Delfi teemalehed). Samuti ei vaadelda „Eesti Ekspressi“ artikleid kui nädalalehe-tüüpi artikleid. Eemaldatakse ka võimalikud topeltlingid.

Iga artikli kohta saadakse *BeautifulSoup* raamistikku kasutades viis andmevälja:

- pealkiri *(title);*
- artikli tekst *(text);*
- väljaande nimi *(subject);*
- publitseerimise kuupäev *(date);*
- artikli link *(link).*

Andmete kogumisel lähtuti põhimõttest analüüsida uudisartikleid või nende osi, mis on avalikult ja tasuta kõigile kättesaadavad. Seetõttu sai tasuliste artiklite puhul analüüsida ainult tasuta kättesaadavat osa.

Andmete kogumine on ajamahukas tegevus, projekti edasiarendusena peaks uurima andmete kogumise ajamahukuse vähendamist. Analüüsiks kasutatavate Uute Uudiste artiklite arvu vähendati, et kolme väljaande artiklite arv oleks samas suurusjärgus.

Andmete kogumiseks kulunud aeg:

- Delfi – aeg 1:34:46 (6598 artiklit);
- Telegram – aeg 1:29:07 (5081 artiklit);
- Uued Uudised – aeg 0:53:19 (6000 artiklit).

Andmed koondati *pandas DataFrame* andmestruktuuri. Andmestikust filtreeriti välja puuduvate andmeväljadega artiklid. Delfi artiklite puhul eemaldati artiklid pehmete uudistega

(meelelahutus, ilmateade, sport jms). Telegrami ja Uute Uudiste linkide hankimisel oli eelnevalt võimalik valida sobivad rubriigid.

Pärast filtreerimist saadud andmestikes on kokku 17 437 uudisartiklit.

Väljaanne	**Artiklite arv**	**Osakaal, %**
Delfi	6363	37%
Telegram	5074	29%
Uued Uudised	6000	34%
Kokku artikleid	17437	

Tabel 2. Uuritavate väljaannete osakaalud.

Andmed salvestatakse csv-failidena:

- data_delfi_UUS.csv;
- data_telegram_UUS.csv;
- data_uueduudised_UUS.csv.

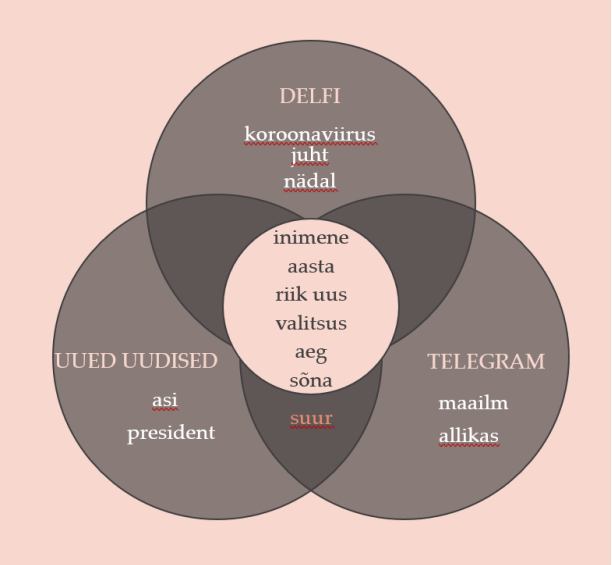
8. Sõnapilved

Sõnapilves esitatakse sagedamini esinevad sõnad suuremas kirjas, harvemini esinevad sõnad on väiksemas kirjas. Sõnapilvede loomisel saab kasutada stoppsõnu. Stoppsõnad on teksti sisu analüüsimisel väikese tähtsusega sõnad, näiteks sidesõnad või asesõnad. Stoppsõnu saab projekti edasises arenduses muuta ja täiendada. Teksti lemmatiseerimiseks kasutatakse EstNLTK vahendeid. Sõna lemma on sõna algvorm (ma-tegevusnimi verbide puhul, ainsuse nimetav nimisõnade ja omadussõnade puhul). Sõnapilve moodustamine on lemmatiseerimise tõttu ajamahukas.

Iga uuritava väljaande jaoks on koostatud kaks sõnapilve, esimeses sõnapilves kajastuvad ainult nimi- ja omadussõnad, teises sõnapilves ainult verbid (Sõnaliikide tabel). Tabelis 3 on lisaks sõnapilvedele toodud väljaannete lõikes kasutatud nimi- ja omadussõnade TOP10 ning verbide TOP10 esinemissageduste ja osakaaludega. Kõik kolm meediaväljaannet kasutavad kõige sagedamini kolme sõna: *inimene, aasta, riik*. Verbide puhul on kõigi väljaannete puhul sarnane kasutada kolme sõna: *olema, ei, saama*. Viimane tulemus ei ole väga informatiivne, sest *olema*-vormide kasutamine on eesti keelele väga omane. Sõna *ei* verbide hulgas tähendab eitavaid verbe: *ei ole, ei olnud, ei saa, ei saanud* jne. Verbe kasutatakse kõigis kolmes meediaväljaandes sarnaselt.

Erinevused TOP10 nimi- ja omadussõnade puhul:

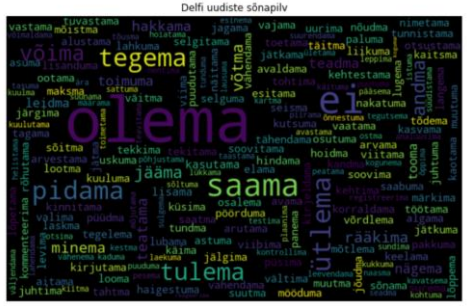
- ainult Delfi puhul on esikümnes sõnad *koroonaviirus, juht, nädal;*
- ainult Telegrami puhul on esikümnes sõnad *maailm, allikas;*
- ainult Uued Uudised puhul on esikümnes sõnad *asi, president.*



Joonis 4. Nimi- ja omadussõnade jaotus TOP 10.

Delfi uudiste nimisõnade ja omadussõnade TOP 10

	sõna	esinemissagedus	osakaal %
9036	inimene	16825	1.984642
897	aasta	7615	0.898250
29445	riik	6558	0.773568
14663	koroonaviirus	5325	0.628126
37803	uus	5301	0.625295
38482	valitsus	5196	0.612909
1128	aeg	5162	0.608899
9933	juht	4709	0.555464
32936	sõna	4699	0.554284
22880	nädal	4154	0.489997



Delfi uudiste verbide TOP 10

	sõna	esinemissagedus	osakaal %
1400	olema	100621	22.888332
196	ei	23805	5.414941
1831	saama	14718	3.347914
2633	ütlema	10291	2.340901
1520	pidama	9534	2.168706
2135	tegema	9226	2.098645
2241	tulema	8249	1.876406
2544	võima	6838	1.555445
66	andma	5103	1.160783
530	jääma	3971	0.903286

Telegrami uudiste nimisõnade ja omadussõnade TOP 10

	sõna	esinemissagedus	osakaal %
15122	inimene	15325	1.632478
1631	aasta	12703	1.353172
46296	riik	5840	0.622099
1961	aeg	4899	0.521860
50885	suur	4653	0.495655
30123	maailm	4502	0.479570
58997	uus	4288	0.456774
60007	valitsus	4203	0.447720
2998	allikas	3966	0.422474
51559	sõna	3774	0.402021

Telegrami uudiste verbide TOP 10

	sõna	esinemissagedus	osakaal %
1778	olema	108854	23.081800
252	ei	26120	5.538580
2302	saama	15259	3.235574
1918	pidama	9181	1.946773
2693	tegema	8855	1.877647
3192	võima	8010	1.698470
3303	ütlema	5651	1.198259
2832	tulema	5375	1.139735
84	andma	5185	1.099446
3218	võtma	4710	0.998726

Uute Uudiste nimisõnade ja omadussõnade TOP 10

	sõna	esinemissagedus	osakaal %
11985	inimene	9287	1.156299
1072	aasta	8683	1.081097
40296	riik	8228	1.024446
52470	valitsus	4596	0.572235
1436	aeg	4322	0.538120
51543	uus	4143	0.515834
44583	suur	3450	0.429550
45145	sõna	3352	0.417348
3494	asi	3081	0.383607
36112	president	2905	0.361694

Uute Uudiste verbide TOP 10

	sõna	esinemissagedus	osakaal %
1835	olema	97165	23.354052
264	ei	27459	6.599896
2380	saama	12442	2.990492
1970	pidama	9342	2.245392
2774	tegema	8004	1.923798
2908	tulema	6689	1.607732
3408	ütlema	6074	1.459914
3288	võima	6028	1.448857
3310	võtma	4441	1.067415
88	andma	4287	1.030400

Tabel 3. Nimi- ja omadussõnade ning verbide sõnapilved väljaannete lõikes. Nimi- ja omadussõnade ning verbide TOP 10 väljaannete lõikes.

9. Andmestike ühendamine

Andmestikud liidetakse failis *03_andmete_liitmine.ipynb*. Andmed segatakse ja salvestatakse faili *data_uudised_koos_UUS.csv*.

Liidetud andmestikus on 17 437 artiklit juhuslikus järjekorras. Iga artikli kohta on viis andmevälja:

- pealkiri *(title);*
- artikli tekst *(text);*
- väljaande nimi *(subject);*
- publitseerimise kuupäev *(date);*
- artikli link *(link).*

10. Meelsuse analüüs

The Global Disinformation Index pakub usaldusväärset ja neutraalset hinnangut uudisportaalide väärinformatsiooni riskide kohta.

Väljaandes „Väärinforiskid Eesti meediaturul“ on teiste väärinfole viitavate näitajate hulgas märgitud artiklite tonaalsust. "Artiklite tonaalsus on hea indikaator, mille abil ennustada uudisportaali teisi väärinforiskidega seotud tugevuste ja nõrkuste seost teiste muutujatega. Siin on oluline välja tuua tähelepanek, et artikli tonaalsuse ja väärinfo esinemise tõenäosuse vahel on tugev seos" (GDI, 2020).

Meelestatuse analüüs (*sentiment analysis*) on meetod, mille abil saab hinnata, kas tekstis kasutatud sõnavara on valdavalt positiivne või negatiivne (Meelestatuse analüüs).

Tekstis esinevad positiivsed ja negatiivsed sõnad loetakse kokku, aluseks võetakse eesti keele uurijate poolt koostatud emotsioonileksikon (Pajupuu, Altrov, Pajupuu, 2016). Emotsioonileksikonis on ligikaudu 39 000 sõna, igale sõnale on määratud vastavalt sõna positiivsusele või negatiivsusele skoor. Negatiivsete sõnade skoor on -1, positiivsete sõnade skoor +1, eriti negatiivsete sõnade skooriks on määratud -8.

Teise meelsuse näitajana võeti kasutusele esimeses isikus olevate sõnade arv. Viiteid eestikeelsete tekstide uurimisel esimese isiku kasutamisele ei leitud. Ingliskeelsete tekstide analüüsimisel on esimese isiku kasutamist uuritud Twitteri-säutsudes (Coppersmith, Dredze, Harman, 2014).

Uudisartiklite meelsuse määramisel kasutati EstNLTK teeki sõnaliikide tuvastamiseks. Artiklite tekstides loendati:

1) esimesele isikule viitavad verbivormid *'gem', 'ks', 'ksime', 'ksin', 'me', 'n', 'neg gem', 'neg ks', 'neg me', 'neg nud', 'neg nuks', 'neg o', 'neg vat', 'nuks','nuksime', 'nuksin', 'nuvat', 'sime', 'sin', 'vat'* (Sõnaliikide tabel);
2) esimesele isikule viitavad asesõnad (*'mina', 'ma', 'meie', 'me'*).

Uudisartiklite meelsusanalüüs on failis *04 Meelsus.ipynb*.

Iga uudisartikli kohta leiti:

1) uudise pikkus sõnades;

2) negatiivsete sõnade arv;

3) esimeses isikus kasutatud sõnade arv;

4) uudise skoor = negatiivsete sõnade arv + esimeses isikus kasutatud sõnade arv;

5) uudise kordaja = uudise skoor/uudise pikkus sõnades;

6) positiivsete sõnade arv.

Kogu korpuse kohta leiti: eelmistes punktides toodu, lisaks artiklites kasutatud negatiivsete ja positiivsete sõnade list. Positiivsete sõnade arv ei kajastu kordajas.

Programmi töö tulemusena väljastatakse uudisartiklite tabel, mille veergudeks on:

1) sõnade arv uudises;
2) esimeses isikus sõnade arv;
3) negatiivsete sõnade arv;
4) uudise kordaja;
5) positiivsete sõnade arv;
6) väljaande nimi;
7) uudise pealkiri.

Uudiste meelsuse analüüsimine on ajamahukas tegevus, 17 437 uudisartikli analüüsimiseks kulus aega 1:26:16.

Meelsuse hindamiseks filtreeriti välja uudised, mille pikkus sõnades on suurem kui 50. Kõrgeima negatiivsuse hinnangu sai rohkete negatiivsete sõnadega Uute Uudiste artikkel ““Valitsusele molli!” Delfi on hakanud kasutama rullnoka kõnepruuki.“

„Delfi/Eesti Päevalehe retoorika muutub päev-päevalt räigemaks. Täna avaldati Vahur Kooritsa arvamuslugu pealkirja all „Kevadine referendum annab rahvale võimaluse valitsusele molli anda“. „Rahvahääletust tagant tõukav EKRE toetub küsitlustulemustele, mille kohaselt toetab enamik Eesti inimesi abielu üksnes mehe ja naise liiduna,“ kirjutab loo autor, lisades, et küsitlus võib muutuda ametlikust teemast selliseks, „kas valijad tahavad valitsusele molli anda“. Sünonüümsõnastiku järgi tähendab “molli andma” eriti labase väljendina vastu hambaid andma, näkku lööma, peksma, kaklema. Nagu näeme, ei käi viha õhutamine meedias enam üksikisiku tasemel ja viisakate väljenditega, vaid ikka banaalselt ja grupiviisiliselt.“ (Uued Uudised, 26.10.2020)

Projekti edasises arendamises võiks täpsustada meelsust leidvat kordajat ning uurida täpsemalt seoseid väärinfo ja artikli tonaalsuse vahel.

11. Uudisartiklite tekstide klasteranalüüs

Uudisartiklite klasterdamise eesmärgiks oli püüda leida artiklite vahelisi sarnasusi, lähtudes artiklite tekstidest. Klasterdamismeetod vajab arvulisi andmeid, seetõttu tekstiliste andmete töötlemiseks peab tekstid vektoriseerima. Vektoriseerimise käigus luuakse iga tekstides esineva sõna jaoks vektor, mille iga positsioon kirjeldab sõna esinemist ühes uudisartiklis. Kõikide sõnade vektoritest pannakse kokku maatriks, mille järgi saab leida nii üksiksõnade esinemismustreid kui ka uudisartiklit iseloomustava sõnavara mustri.

Uudisartiklite klasteranalüüsiks kasutati uudiste tekstide vektoriseerijana *TfidfVectorizer*-it. TF-IDF kasutab sõnasageduste asemel skoori, mis näitab, kuivõrd iseloomulik on sõna antud uudisartiklile. TF (*term frequency*) on sõna esinemissagedus tekstis ja IDF (*inverse document frequency*) kajastab sõna esinemissagedust kõigis tekstides. TF-IDF on nimetatud skooride korrutis.

Vektoriseerijat kasutades saab muuta mitmesuguseid parameetreid ja kasutada erinevaid eeltöötluse võtteid. Tekstide klasterdamine on failis *05 Klasterdamine.ipynb*.

Projektis kasutati vektoriseerijat järgnevalt:

```
tfidf_vectorizer = TfidfVectorizer(
                    max_df=0.8,
                    max_features=200000,
                    min_df=0.2,
                    stop_words=punc,
```

use_idf=True,

tokenizer = lemmatize_with_estnltk,

ngram_range=(1,3))

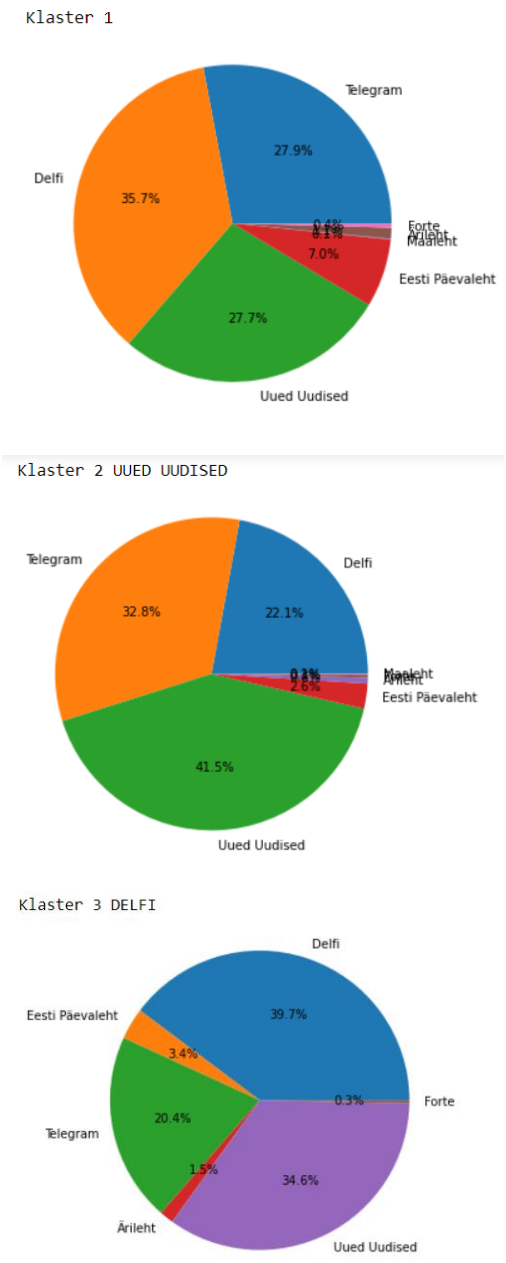
Kasutatud parameetrid:

- *max_df=0.*8 analüüsiti sõnu, mille esinemissagedus on kuni 80%;
- *min_df=0.2* analüüsiti sõnu, mille esinemissagedus on alates 20%;
- *max_df* ja *min_df* kasutamise eesmärgiks on jätta analüüsist välja liiga sagedased ja liiga harva kasutatud sõnad;
- *max_features=200000* maksimaalne sõnade hulk;
- *stop_words=punc* stoppsõnadena kasutati kirjavahemärke ja teisi sümboleid;
- *punc = ['.', ',', '"', "'", '?', '!', ':', ';', '(', ')', '[', ']', '{', '}',"%", '‘', '’', ',,', '-', '’’];*
- *tokenizer = lemmatize_with_estnltk* sõnad lemmatiseeriti EstNLTK vahenditega;
- *ngram_range=(1,3)* analüüsiks kasutati sõnade n-gramme (vaadeldi sõnu ühekaupa, kahekaupa ja kolmekaupa).

Klastrite arvuks valiti kolm, klasterdamismeetodina kasutati k-keskmiste meetodit. Ajakulu vähendamiseks kasutati klasterdamisel 10 000 juhuslikult segatud uudisartiklit. Joonisel 5 on esitatud uudisartiklite jaotumine kolme klastri vahel.

Joonis 5. Uudisartiklite jaotumine kolme klastrisse.

Järgnevalt uuriti kuidas erinevad väljaanded jaotuvad erinevate klastrite vahel. Üheski klastris ei tekkinud mõne väljaande tuntavat ülekaalu. Kõige tihedamalt on koondunud kolmas klaster.



Joonis 6. Väljaandjate jaotumine klastrite vahel.

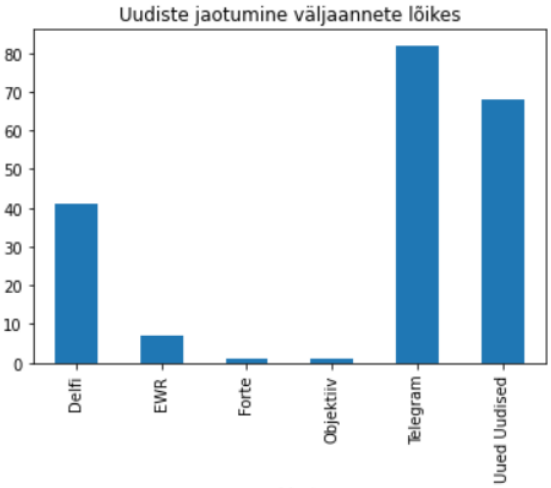
Artiklite jaotumisel klastritesse võib märgata, et klastrisse 2 on koondunud üle 40% Uute Uudiste artikli, klastrisse 3 on koondunud ligi 40% Delfi uudistest. Klastritele sisulise tähenduse andmine on projekti selles etapis veel komplitseeritud ning nõuab täiendavat uurimist ja selgitamist. Samuti vajaksid katsetamist klasteranalüüsi erinevad parameetrid.

Täiendavalt uuriti ka hierarhilist klasterdamist. Hierarhiline klasterdamine on sobilik kasutada pigem väiksemamahulise andmestiku korral.

Klasteranalüüsimisel on aluseks võetud Brandon Rose kirjutatud koodi (Brandon Rose klasteranalüüs).

12. Väärinfo ja tõese info uurimine

Vaadeldakse uudisartikleid, mis on jagatud väärinfoks ja tõeseks infoks. Projekti autoril ei olnud kasutada valmisandmestikku vale ja tõese infoga. Autor valis välja umbes 2000 artikli hulgast 200 artiklit, mis autori subjektiivsel hinnangul märgistati tõeseks infoks ja valeinfoks. Lisaks väljaannetele Delfi, Telegram, Uued Uudised, kaastati uurimisse artiklid Objektiivi ja EWR lehekülgedelt (Delfi arhiiv, Telegram arhiiv, Uued Uudised arhiiv, Objektiiv, EWR).



Delfi	41
EWR	7
Forte	1
Objektiiv	1
Telegram	82
Uued Uudised	68

Joonis 7. Väärinfo uurimisega haaratud väljaanded.

Eestis tegutseva Ekspress Meedia juurde kuuluva Faktikontrolli põhimõtted on:

"**Meie tööd iseloomustab faktipõhisus.** Enne faktiväite kontrollima asumist veendutakse, et tegemist pole arvamusega. Seejärel tehakse taustauuringud, et kontrollitav fakt avalikult kättesaadavate kirjalike või suuliste materjalide põhjal ümber lükata või kinnitada. Suulised allikad on kontrollitavas valdkonnas hinnatud spetsialistid. Kirjalikud allikad võivad olla nii eesti- kui ka võõrkeelsed teadustööd, aga ka usaldusväärsete meediakanalite populaarteaduslikud artiklid. Võimalusel toetutakse faktikontrollis alati enam kui ühele allikale. Juhul, kui eksitavaid väiteid on kontrollitavas sisus rohkem kui üks, eraldatakse need arusaadavalt.

Meie tööd iseloomustab läbipaistvus. See tähendab nii seda, et meie allikad, toimetajad kui ka peamised rahastusallikad on avalikud. Meie tagasiside vorm on kättesaadav kõigile huvilistele. Faktikontrollide hilisemal täpsustamisel või parandamisel tuuakse muudatused eraldi välja. Võimalusel võetakse ühendust isikuga, kelle osas faktikontrolli läbi viiakse.

Meie tööd iseloomustab erapooletus. Me ei kontrolli vaid ühe kitsa ühiskonnagrupi ega vaid ühe poliitilise erakonna või liikumise väiteid. Jälgime kõikide gruppide avaldusi ning nende ulatust ja mõju. Enne kontrolli alustamist teeme lisaks taustatööle ülevaate senistest kontrollidest. Iga faktikontroll läbib toimetamisprotsessi. See tagab, et faktikontrolli hinnang ei sõltu vaid ühest inimesest. Meie autorid väldivad poliitilisi seisukohti." (Ekspress Meedia Faktikontroll).

Projekti autor üksinda eelpooltoodud tingimusi täita ei suuda. Faktide kontrollimisega tegelevad meeskonnad, kaasatakse sõltumatuid eksperte, valeks tunnistatud väite esitajal on õigus vaidele.

Analüüsimiseks peaks kasutama faktikontrolli ja/või sõltumatute ekspertide poolt kontrollitud artikleid. Seetõttu peab praktikaaruande väärinfot kajastava osa tulemusi käsitlema ettevaatusega. Saadud tulemused võivad olla sõltuvuses projekti autori subjektiivsetest hinnangutest.

Väärinfo ja tõese info uurimise väärtuslikum osa on tehniline külg, antud projektis uuriti milliseid võimalusi EstNLTK raamistik tekstiliste andmete uurimisel pakub.

Tõese info ja valeinfo uurimisel saadud tulemused on failis

06 Väärinfo_ja_tõese_info_uurimine.ipynb.

Iga uudisartikli kohta on kuus andmevälja:

1) väljaande nimi;

2) artikli pealkiri;

3) artikli tekst;

4) artikli publitseerimise kuupäev;

5) artikli link;

6) tõene või väär (*fake*, *true*).

Väärinfo sõnapilv

Tõeste uudiste sõnapilv

Väärinfo nimisõnade ja omadussõnade TOP 10

	sõna	esinemissagedus	osakaal %
1260	inimene	349	1.670816
71	aasta	233	1.115473
4604	test	131	0.627154
2554	maa	129	0.617579
4242	suur	111	0.531406
5309	viirus	111	0.531406
103	aeg	108	0.517043
5114	valimine	94	0.450019
5446	võimalik	92	0.440444
4745	tulemus	92	0.440444

Tõeste uudiste nimisõnade ja omadussõnade TOP 10

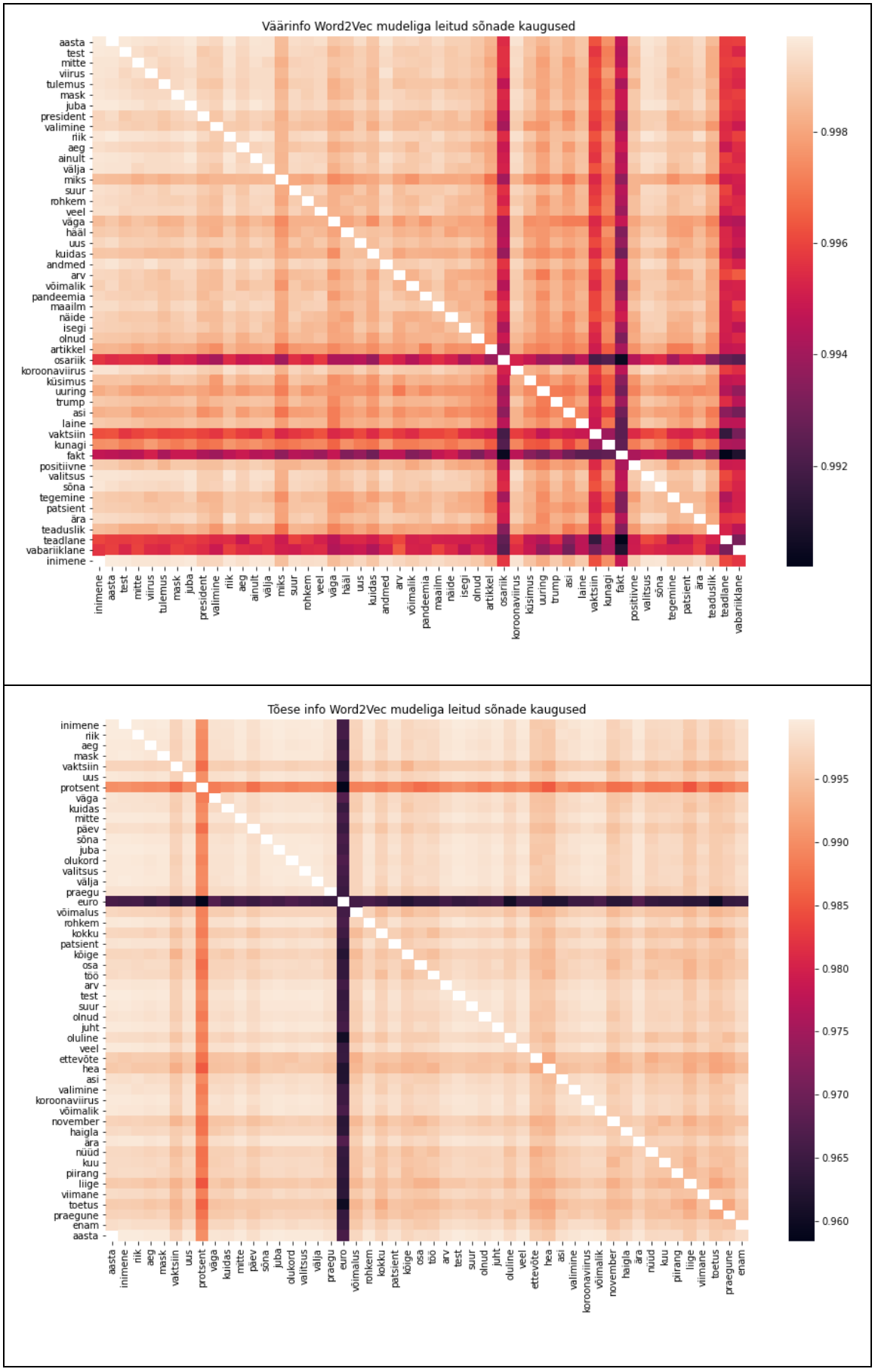
	sõna	esinemissagedus	osakaal %
40	aasta	197	1.482764
833	inimene	181	1.362336
63	aeg	119	0.895680
2822	riik	110	0.827939
2606	päev	77	0.579557
3722	uus	76	0.572031
1645	laps	76	0.572031
1913	mask	74	0.556977
3631	töö	64	0.481710
3764	vaktsiin	64	0.481710

Tabel 4. Väärinfo ja tõese info nimi- ja omadussõnade kasutuse võrdlus.

Väärinfot sisaldavates uudisartiklites oli kõige sagedasem sõna *inimene*, tõese infoga artiklites sõna *aasta*. Kui väärinfot sisalduvates artiklites oli sagedasti juttu *viirusest*, siis tõest infot sisalduvates tekstides räägiti *maskidest* ja *vaktsiinist*. Ka *lastest* räägiti tõestes uudistes rohkem.

Word2Vec on tehisnärvivõrkudel põhinev keelemudel. Word2Vec meetodiga esitatakse sõnad vektoritena. Sisendkorpuse sõnadega leitakse korpuses esinevatele sõnadele vektorid, kus saadud vektorid rühmitavad sarnaseid sõnu. Kahe sõna vaheline sarnasus leitakse neile vastavate vektoritevahelise koosinuskaugusega, mistõttu jääb sarnasust määrav arv vahemikku -1 kuni 1 (kõige sarnasem) (Tartu Ülikooli õppeaine Eesti keele töötlus Pythonis materjal).

Järgnevas tabelis 5 on visualiseeritud tekstides esinevate 50 kõige sagedasemate sõnade sarnasused. Mida sarnasem sõna teisele sõnale, seda lähemal on väärtus arvule 1. Väärinfot sisaldavate artiklite puhul võib märgata, et sõnade vahelised sarnasused on väiksemad võrreldes tõest infot sisaldavate artiklitega. Uuritava korpuse väiksuse tõttu ning valikusse sattunud tekstide subjektiivse tõese-väära jaotuse tõttu ei tohiks siinkohal teha põhjalikke järeldusi. Kuid saadud tulemus on huvitav ning tasuks kaaluda edasist uurimist.



Tabel 5. Word2Vec mudelid tõese info ja väärinfo jaoks.

13. Väärinfo mudelid

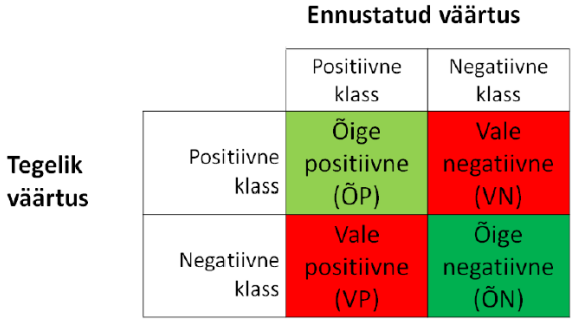
„Masinõpe tegeleb matemaatiliste (statistiliste) mudelitega, mida on võimalik treenida empiiriliste andmetega nii, et need mudelid suudavad teha piisavalt täpseid ennustusi või otsuseid ka sama tüüpi uute andmete korral. Enamik mudeleid sisaldavad vabu parameetreid, mida optimeeritakse treenimise käigus, püüdes minimeerida mõnesugust sihi- või kahjufunktsiooni (*objective function*, *loss function*).“ (Kiisk, V).

Eesmärk on leida juhendatud klassifitseerimisalgoritmidega mudel, mis oskaks ennustada, kas etteantud uudis on tõene või vale. Andmed (200 uudisartiklit) jagatakse treenimisandmestikuks (70%) ja testandmestikuks (30%).

Klass *CountVectorizer* võimaldab luua sõnavektoreid, mille iga positsioon kirjeldab sõna esinemissagedust ühes dokumendis. Klassi initsialiseerimisel on võimalik täpsustada mitmeid teksti eeltöötluse samme, nt kas sõnad tuleks teisendada väiketähelisteks, kuidas jagada tekst sõnadeks ning millised sõnad on stoppsõnad, mis tuleks välja jätta.

TfidfVectorizer kasutab sõnasageduste asemel TF-IDF skoori, mis näitab, kuivõrd iseloomulik on sõna mingile tekstile korpuses. TF (*term frequency*) on sõna esinemissagedus tekstis, IDF (*inverse document frequency*) kajastab seda, kui paljudes korpuse dokumentides see sõna üldse esineb, TF-IDF on nende korrutis (Tartu Ülikooli õppeaine Eesti keele töötlus Pythonis materjal).

Veamaatriks, ka segadusmaatriks, eksimismaatriks (*confusion matrix*) väljendab saadud mudeli headust. Maatriksi peadiagonaalil asetsevad kõik klassifikaatori pool õigesti ennustatud väärtused: õiged positiivsed, õiged negatiivsed. Kõik, mis jääb peadiagonaalilt välja, on väärad: vale positiivsed, vale negatiivsed. Mida rohkem õigeid negatiivseid ja õigeid positiivseid tulemusi, seda parem mudel.



Joonis 8. Veamaatriks.

Mudeli headust hinnatakse täpsusmäära (*accuracy*) abil. Täpsusmäär näitab, milline oli õigete ennustuste osakaal kõikidest ennustustest.

$$täpsusmäär = \frac{õige\ pos + õige\ neg}{õige\ pos + õige\ neg + vale\ pos + vale\ neg}$$

Pipeline'i on ühendatud eeltöötluse sammud: vektoriseerimine, teisendamine ja klassifitseerimine. Mudelite katsetamisel saab muuta sõnestajat, stoppsõnade loetelu ja teisi parameetreid.

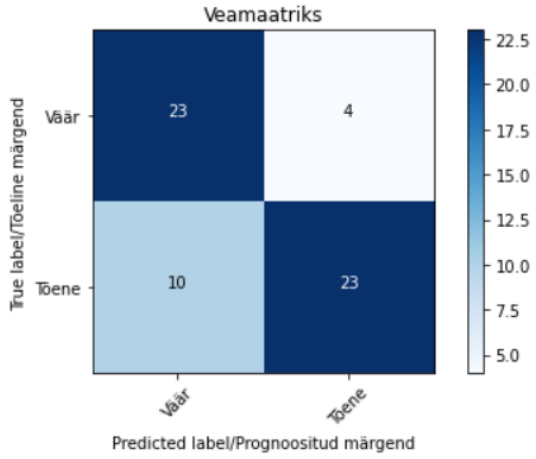
Andmete peal katsetati nelja klassifitseerimismeetodit.

13.1 Logistiline mudel *(Logistic regression)*

Logistilist regressiooni kasutatakse kaheklassilise diskreetse väärtuse ennustamiseks:

1) eeltöötluseks kasutati tekstide lemmatiseerimist *tokenizer = lemmatize_with_estnltk*;
2) stoppsõnadena kasutati asesõnu ja sidesõnu;
3) sõned vektoriseeriti *CountVectorizer*;
4) teisendati TF-IDF skooriks *TfidfTransformer*;
5) klassifitseerijana kasutati *LogisticRegression.*

Logistilise regressiooni mudeli ennustustäpsuseks saadi 76,67%.



Joonis 9. Logistilise regressiooni veamaatriks.

Mudel annab kümme valenegatiivset ja neli valepositiivset tulemust.

13.2 Otsustuspuu *(Decision Tree)*

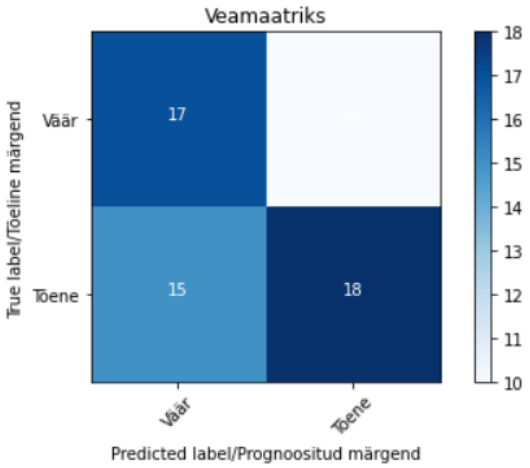
Otsustuspuu on traditsiooniline osa andmekaevest ja masinõppe algoritmidest. Otsustuspuu ei ole alati parima täpsusega. Otsustuspuu kasutab struktuuri, kus aluseks on üks sõlm, mis jaguneb erinevateks harudeks (oksteks), mis omakorda tipnevad sõlmedega, millest igaüks

võib edasi hargneda või lõppeda lehega. Iga sõlme juures on test või küsimus, mis määrab, millist haru mööda edasi minna kuni jõutakse leheni ehk otsuseni (Masinõppe algoritmid).

Otsustuspuu meetodi kasutati järgevalt:

1) eeltöötluseks kasutati tekstide lemmatiseerimist *tokenizer = lemmatize_with_estnltk*;
2) stoppsõnadena kasutati asesõnu ja sidesõnu;
3) sõned vektoriseeriti *CountVectorizer*;
4) teisendati TF-IDF skooriks *TfidfTransformer*;
5) klassifitseerijana kasutati *DecisionTreeClassifier* (criterion= 'entropy', max_depth = 2, splitter='best', random_state=42).

Otsustuspuu mudeli ennustustäpsuseks on 58,33%.



Joonis 10. Otsustuspuu veamaatriks.

Mudel eksis 15 korral, saades 15 valenegatiivset tulemust.

13.3 Juhuslik mets *(Random Forest)*

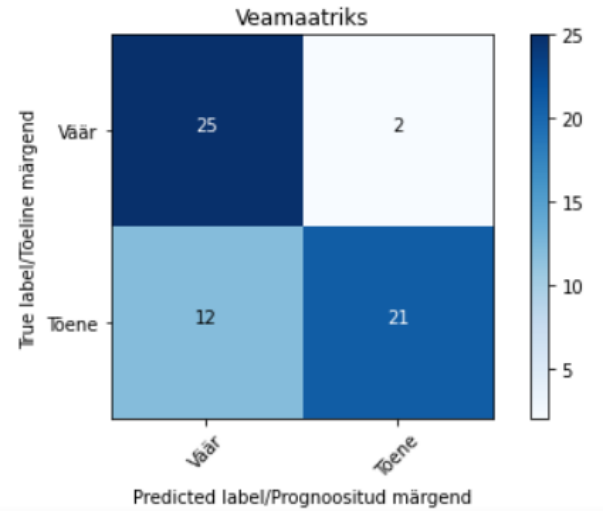
Mitme mudeli otsustuspuu kombineerimine ühte mudelite kogumisse annab otsustuspuude metsa. Juhusliku metsa puhul kasutatakse juhuslikku valikut nii vaatluste kui parameetrite valikul. Selle tulemusena on võimalik saavutada suurem sõltumatus andmete muutumisest, andmetes sisalduvast mürast ja ekstreemsetest vaatlustest ning ülemäärasest sobitamisest algandmetele. Samuti on juhusliku metsa eeliseks parem toimetulek tasakaalustamata treeningandmetega (Masinõppe algoritmid).

Juhuslik mets ei nõua andmete eeltöötlemist, sest andmeid ei pea normaliseerima. Samuti pole vaja tegeleda parameetrite valikuga, sest algoritm teeb seda ise. (Masinõppe algoritmid).

Juhusliku metsa meetodit kasutati järgevalt:

1) mudeli omaduste tõttu eeltöötlust ja stoppsõnu ei kasutatud;
2) sõned vektoriseeriti *CountVectorizer*;
3) teisendati TF-IDF skooriks *TfidfTransformer*;
4) klassifitseerijana kasutati *RandomForestClassifier* (n_estimators=50, criterion="entropy").

Juhusliku metsa mudeli täpsus on 76,67%.



Joonis 11. Juhusliku metsa veamaatriks.

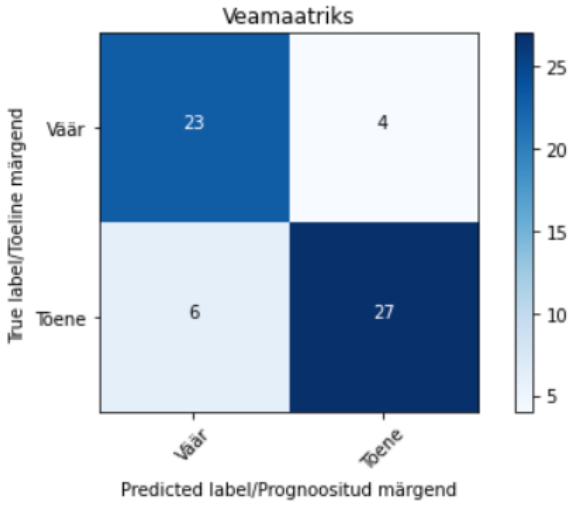
Mudel eksis 14 korral, saadi 12 valenegatiivset ja kaks valepositiivset tulemust.

13.4 Passiiv-agressiiv mudel (*Passive Agressive*)

Passiiv-agressiiv mudelit kasutati järgenvalt:

1) eeltöötluseks kasutati tekstide lemmatiseerimist *tokenizer = lemmatize_with_estnltk*;
2) stoppsõnadena kasutati asesõnu ja sidesõnu;
3) sõned vektoriseeriti *TfidfVectorizer*;
4) teisendati TF-IDF skooriks *TfidfTransformer*;
5) klassifitseerijana kasutati *PassiveAggressiveClassifier*(max_iter=50).

Passiiv-agressiiv mudeli täpsuseks saadi 83,33%.



Joonis 12. Passiiv-agressiiv mudeli veamaatriks.

Passiiv-agressiiv mudel eksis kümnel korral, saadi kuus valenegatiivset ja neli valepositiivset tulemust. Passiiv-agressiiv mudeli täpsus osutus uuritud neljast mudelist parimaks.

Klassifitseerimismudelite kvaliteeti saab hinnata ristvalideerimisega. Tekstid jagatakse juhuslikult k erineva, kuid sama suurusjärguga osa vahel (k on kasutaja poolt määratud parameeter). Ristvalideerimismeetodis kasutatakse üht osa tekstidest täpselt ühe korra testandmetena ja k-1 korda treeningandmetena (Vaik, Muischnek).

Kümnekordse ristvalideerimisega (k väärtuseks valiti 10) saadi passiiv-agressiiv mudeli keskmiseks täpsuseks 85,5%. Seega parima täpsuse andis 10-kordse ristvalideerimisega passiiv-agressiiv mudel.

Praktikaprojekti kokkuvõte

Praktika peamised ülesanded olid:

- saada ülevaade väärinfo levitamisest kui ühest infoturbe probleemist;
  Praktika käigus on uuritud väärinfo levitamise ja levimise trende, saadud hea ülevaade käesoleval hetkel levivatest väärinfo trendidest kirjutavas eestikeelses meedias. On tundma õpitud väärinfo levitamise vastu tegutsevaid organisatsioone (nt GDI, Ekspress Meedia Faktikontroll, Euroopa Liidu Infokeskus). Tutvuti erinevate meediaväljaannete artiklitega.
- arendada algoritmilist mõtlemist ja süvendada programmeerimise oskusi;
  Praktika projekti programmeerimiskeeleks oli Python. Pythoni koodi rakendati *Jupyter Notebook* keskkonnas. Tekstide töötluseks kasutati EstNLTK raamistiku. Projekti autor oli eelnevalt kokku puutunud programmeerimiskeelega Python. Selle projekti raames õppis

projekti autor kasutama Jupyter Notebook keskkonda, kasutama EstNLTK raamistikku ning laiendas teadmisi programmeerimiskeelest Python.

- katsetada väärinfo ja tõese info põhjal tekstianalüüsi kasutades EstNLTK teegi võimalusi eestikeelsete tekstide töötlemiseks;

  EstNLTK teek on spetsiaalselt kohandatud eesti keele eripäradele ja vajadustele. Ingliskeelses kirjanduses on rohkelt näiteid väärinfo klassifitseerimiseks ja klasterdamiseks. Eestikeelsete tekstide jaoks on tarvis lisaks tunda EstNLTK teegi omadusi ja võimalusi, aega kulus erinevatele katsetustele, praktikaprojekti kõik katsetused ei jõudnud.
- luua detektor väärinfo tuvastamiseks;

  Praktikaprojekti viimases osas keskenduti väärinfo tuvastamisele juhendatud masinõppe algoritmidele toetudes. Õpiti tundma erinevaid klassifitseerimismeetodeid. Koostati neli mudelit tõese info ja väärinfo eristamiseks. Hinnati klassifitseerimismudeli kvaliteeti ristvalideerimisega.

Kokkuvõttes saab öelda, et praktikaprojekti eesmärgid üldjoontes täideti. Järgmine samm võiks olla mõne uuritud teemaga süvitsi minek.

Tänuavaldus

Suur tänu minu juhendajale, Jan Willemsonile, huvitava teema leidmise ja võimaluse eest tutvuda minu jaoks täiesti uue teemavaldkonnaga. Soovin tänada Cybernetica AS, kes võimaldas mulle läbida praktikat ja tutvuda ettevõtte tegevustega.

Lõpetuseks

```
In [172]:
txt1 = ['Maa on lame!']
pac.predict(tfidf_vectorizer.transform(txt1))[0]

Out[172]: 'fake'
```

Kasutatud kirjandus

1. **Cooke**, N.A. (2017). Posttruth, Truthiness, and Alternative Facts: Information Behavior and Critical Information Consumption for a New Age. Library Quarterly: Information, Community, Policy, 87, (3), 211-221. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
2. **Coppersmith**, Glen; **Dredze**, Mark, **Harman**, Graig (2014). Quantifying Mental Health Signals in Twitter, Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 51–60, Baltimore, Maryland USA, June 27, 2014.
3. **Hennoste**, T. (2008). Uudise käsiraamat. Tartu: Tartu Ülikooli kirjastus.
4. **Himma-Kadakas**, Marju (2017). "Alternative facts and fake news entering journalistic content production cycle". *Cosmopolitan Civil Societies: an interdisciplinary journal*.
5. **Klaassen, Maia**. (2018). Vale- ja võltsuudiste avaldamine peavoolumeedias: Eesti meediaväljaannete peatoimetajate selgitused tekkepõhjustele. Bakalaureusetöö. https://dspace.ut.ee/bitstream/handle/10062/60632/klaassen_maia_ba_2018.pdf?sequence=1&isAllowed=y
6. **Pajupuu**, Hille; **Altrov**, Rene; **Pajupuu**, Jaan (2016). Identifying polarity in different text types. Folklore. Electronic Journal of Folklore, 64, 25−42. DOI PDF
7. **Vaik**, Kristiina; **Muischnek**, Kadri. Eestikeelsete veebitekstide automaatne liigitamine, Eesti rakenduslingvistika ühingu aastaraamat, 14, 215–227, https://www.researchgate.net/publication/324801038_Eestikeelsete_veebitekstide_automaatne_liigitamine/fulltext/5ae32f0d458515c60f68b2f4/Eestikeelsete-veebitekstide-automaatne-liigitamine.pdf
8. **Wardle**, C., & **Derakhshan**, H. (2017). Information Disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe Report, 27. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
9. Tartu Ülikooli õppeaine **„Eesti keele töötlus Pythonis“** materjal, 2020.

Internetiallikad

1. **Euroopa Liidu Infokeskus**, (viimati vaadatud 15.12.2020), https://elik.nlib.ee/et/fookuses/desinformatsioon
2. **Ekspress Meedia Faktikontroll**, (viimati vaadatud 02.01.2021) https://epl.delfi.ee/artikkel/91897223/meie-pohimotted
3. **Eesti Päevaleht, 31.12.2020**. „TOP 5 | Milliseid valesid jäid Eesti inimesed tänavu uskuma? Maskivalede ning ebaravi lubaduste õnge langeti ohtralt„. (viimati vaadatud 02.01.2021), https://tv.delfi.ee/uudised/paevauudised/top-5-milliseid-valesid-jaid-eesti-inimesed-tanavu-uskuma-maskivalede-ning-ebaravi-lubaduste-onge-langeti-ohtralt?id=92139851

4. **Eesti Ekspress, 29.04.2020**. FOTOD | Tallinna teletornist näeb Soomet paremini kui varem, (viimati vaadatud 15.12.2020), https://ekspress.delfi.ee/artikkel/89694777/fotod-tallinna-teletornist-naeb-soomet-paremini-kui-varem
5. **Telegram, 31.05.2020**. Kas koroonaviirus tõestab, et Maa on lame? , (viimati vaadatud 15.12.2020), https://www.telegram.ee/nwo/kas-koroonaviirus-toestab-et-maa-on-lame
6. **Delfi arhiiv**, (viimati vaadatud 15.12.2020), https://www.delfi.ee/archive/
7. **Telegrami arhiiv**, (viimati vaadatud 15.12.2020), https://www.telegram.ee/arhiiv
8. **Uued Uudised arhiiv**, (viimati vaadatud 15.12.2020), https://uueduudised.ee/arhiiv/
9. **Sõnaliikide tabel**, (viimati vaadatud 31.12.2020), https://github.com/estnltk/estnltk/blob/version_1.6/tutorials/nlp_pipeline/A_02_morphology_tables.ipynb.
10. **GDI 2020**, Väärinfo riskid Eesti media turul, (viimati vaadatud 15.12.2020) https://disinformationindex.org/wp-content/uploads/2020/10/GDI-RiskRatingsReport_Estonian.pdf
11. **Meelestatuse analüüs**, (viimati vaadatud 15.12.2020), http://samm.ut.ee/meelestatuse-anal%C3%BC%C3%BCs
12. **Uued Uudised, 26.10.2020**. "Valitsusele molli!" Delfi on hakanud kasutama rullnoka kõnepruuki, (viimati vaadatud 15.12.2020), https://uueduudised.ee/arvamus/repliik/delfi-retoorika-valitsusele-molli/
13. **Brandon Rose klasteranalüüs**, (viimati vaadatud 31.12.2020) http://brandonrose.org
14. **Masinõppe algoritmid**, (viimati vaadatud 02.01.2021), https://masinope.ee/masinoppimine/
15. **Objektiiv**, (viimati vaadatud 02.01.2021), https://objektiiv.ee/
16. **EWR**, (viimati vaadatud 02.01.2021), https://www.eesti.ca/
17. **Kiisk, Valter**, (viimati vaadatud 02.01.2021), http://kodu.ut.ee/~kiisk/python/machine.html#Klassifitseerimine