doi:10.5128/ERYa14.13

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 14, 215–227

EESTIKEELSETE VEEBITEKSTIDE AUTOMAATNE LIIGITAMINE

Kristiina Vaik, Kadri Muischnek

Ülevaade. Internet on oluline keeleressurss, mille üheks keeleteaduslikuks ja keeletehnoloogiliseks kasutusvõimaluseks on seal leiduvate tekstide koondamine keelekorpuseks. Kuid täisautomaatselt korjatud korpusega seistakse uudse situatsiooni ees: olemas on palju andmeid, ent pole täpselt teada, millist keelematerjali need sisaldavad. Loomuliku keele uurimise ja töötlemise seisukohalt on vajalik tekstide eristamine tekstiliigiti, sest sellest sõltub sobivate töötlusvahendite valik. Artiklis kirjeldame tekstiliikide eristamise lihtsustatud versiooni: korpuse Estonian Web 2013 (etTenTen13) binaarse klassifitseerimise katset, mille eesmärk oli liigitada tekstid kirjakeele normi järgivateks ja mittejärgivateks. Treeningandmetes kasutasime kirjakeele esindajana Tasakaalus korpust ja kirjakeele normi mitte järgivate tekstide esindajana Uue meedia korpust ning testandmetena käsitsi liigitatud Estonian Web 2013 alamkorpust. Klassifitseerimismudelite loomisel rakendasime erinevaid juhendatud masinõppe algoritme ning tunnustena sõnehulkasid. Klassifitseerimismudelite kvaliteeti hindasime 10-kordse ristvalideerimise teel, kus parima tulemuse andis tehisnärvivõrkudel põhinev algoritm, mis 99% täpsusega liigitas dokumendi õigesse klassi. Seejärel katsetasime mudeleid käsitsi liigitatud Estonian Web 2013 testkorpusel, kus parima tulemuse andis taas tehisnärvivõrkudel põhinev algoritm täpsusega 74%.

Võtmesõnad: korpuslingvistika, automaatne liigitamine, klassifitseerimine, keeletöötlus, masinõpe, tekstiliik, keelekorpus, eesti keel

1. Sissejuhatus

Interneti tulek on suurendanud kirjalike tekstide variatiivsust: veebi kolinud klassikaliste kirjalike tekstide kõrvale on tekkinud nn kasutaja loodud sisu – toimetamata, kirjakeele normist hälbivad ning erineva spontaansuse astmega tekstid. Veebi keelekasutus on pakkunud keeleteadlastele uut huvitavat uurimismaterjali ja

arvutilingvistidele uusi huvitavaid ülesandeid, mille hulka kuulub ka tekstiliikide ehk žanrite automaatne liigitamine/klassifitseerimine (ingl *automatic genre identification* või *automatic genre classification*). Info keelenäidete tekstiliigilise kuuluvuse kohta on keeleuurijatele oluline, kuna see annab võimaluse analüüsida erinevate keeleliste väljendusvahendite (mitte)kasutamist tekstiliigiti. Loomuliku keele uurimise ja töötlemise (nt morfoloogiline ja süntaktiline analüüs, masintõlge, infootsing jm) seisukohalt on samuti vajalik, et tekstid oleksid tekstiliigi järgi eristatavad, sest sellest sõltub õige profiiliga töötlusvahendite valik. Näiteks süntaksianalüsaator ei saa eri tüüpi (nt perioodika *vs.* foorumitekstid) tekstiliike analüüsida sama kvaliteediga (vt Särg 2015), sest analüsaatoreid arendatakse üldjuhul kirjakeele normi järgivate tekstide jaoks.

Selles artiklis kirjeldame eestikeelsete tekstide automaatse liigitamise katseid, mis on läbi viidud Estonian Web 2013 alamkorpusega. Estonian Web 2013 (Kallas jt 2015) kuulub nn kolmanda põlvkonna, st veebist korjatud korpuste hulka; selline korpuseloomise viis on tänapäeval tavaline (Jakubíček jt 2013). Veebist korjatud korpuste suur maht ning nendes sisalduvate traditsiooniliste ja uute tekstiliikide arvukus on andnud tõuke nende automaatsele liigitamisele ehk klassifitseerimisele (nt Santini 2007). Veebist korjatud korpuse ammendav klassifitseerimine on keeruline ülesanne, sest juhendatud masinõppe meetodite kasutamine eeldab eelnevat tekstiliikide klassifikatsiooni väljatöötamist ja selle klassifikatsiooni järgi märgendatud treeningkorpuse olemasolu.

Siinkirjeldatud töö eesmärgiks on eelkõige lahendada praktiline ja keeletehnoloogilisest vajadusest lähtuv ülesanne: luua tööriist eestikeelsete veebitekstide liigitamiseks kirjakeele normi järgivateks ja mittejärgivateks. Selline liigitus võimaldab rakendada sobivaid teksti tükeldamise (lõigud, laused, osalaused, sõnad) võtteid, vajadusel kasutada teksti normaliseerimist ning rakendada sobivat (esmase) morfoloogilise analüüsi viisi. Ülesande lahendamiseks kasutame juhendatud masinõppel põhinevaid klassifitseerijaid.

Artiklis käsitleme kõigepealt lühidalt tekstiliigi mõistet ning selle automaatse klassifitseerimise võimalusi. Seejärel tutvustame kasutatud keelekorpusi ning anname ülevaate erinevate klassifitseerimismudelite kvaliteedist treening- ja käsitsi liigitatud testkorpuse liigitamisel. Lõpuks arutleme edasiste töösuundade üle, mis võimaldaksid klassifitseerimise kvaliteeti tõsta.

2. Taust: tekstiliigid ja nende automaatne liigitamine

Artiklis anname ülevaate veebist korjatud korpuse binaarsest liigitusest, kuid suuremas plaanis soovime jõuda mitmeklassilise liigituseni, st klassifitseerida tekste nende tekstiliigi järgi, mistõttu on oluline selgitada tekstide automaatsel klassifitseerimisel kasutatavaid mõisteid *tekstiliik* ja *tekstitüüp*.

Reet Kasik (2007: 35) defineerib tekstiliiki ehk žanrit kui kultuurisidusat keelekasutusviisi, mis on aja jooksul välja kujunenud. Tekstiliigi järgi eristatakse argi-, ilukirjandus- ja tarbekeelt – viimase alaliikidena ajakirjandus-, ameti- ja teaduskeelt, millest igaühel on omakorda alaliike. Tekstitüüp on inimese sihipärane keelekasutamine kindla eesmärgi saavutamiseks ning erinevalt tekstiliikidest on tekstitüüpi defineerivad keelenähtused universaalsed ning selle järgi eristatakse deskriptiivseid, narratiivseid ja argumenteerivaid tekste.

Ingliskeelses kirjanduses kasutatakse ka mõistet *register*, mille puhul sõltuvad sagedasti kasutatavad keelenähtused teksti kommunikatiivsest eesmärgist (Biber, Conrad 2009: 2). Olukorra teeb keeruliseks see, et erialakirjanduses kasutatakse neid termineid läbisegi ja samatähenduslikuna, mis annab autorite terminikasutust kõrvutades keeruka pildi (vt Hennoste 2000: 15–18). Tekstiliikide automaatse klassifitseerimise seisukohalt pole vahet, kas tegu on registri, tekstiliigi või -tüübiga (tunnustena saab vaadelda vaid keelelisi ja vormilisi nähtuseid), kuid selguse mõttes kasutame edaspidi terminit *tekstiliik*.

Selles artiklis nimetame veebitekstideks kõiki internetis leiduvaid tekste. Arvutilingvistikas on interneti nn kasutaja loodud sisu automaatanalüüs, eriti selle jaoks töötlusvahendite kohandamine või teksti normaliseerimine (st tänapäeva kirjakeele normi järgivaks tegemine) aktuaalne teema. Näiteks toimub alates 2015. aastast igal aastal seminar nimega *Workshop on Noisy User-Generated Text*[1], aga sellest, kuidas automaatselt tuvastada, kas tekst vajab sellist eeltöötlust või mitte, on vähem kirjutatud. Ometi on see tänapäeva veebist täisautomaatselt kogutavate keelekorpuste ajajärgul reaalne probleem.

Tekstide automaatse klassifitseerimisega (sh tekstitüpoloogiaga) on tegeletud alates 1980-ndate lõpust. Suurt mõju sellele valdkonnale on avaldanud Douglas Biberi multidimensionaalse analüüsi meetod (1988, 1995 ja 2009), mis kasutab registrite uurimiseks mitmemõõtmelisi statistilisi tehnikaid (peamiselt faktor- ja klasteranalüüs). Tema idee järgi ei tule tekstide liigitamisel lähtuda nende sisulisest ega funktsionaalsest tüübist, vaid iseloomulike keeleliste tunnuste koosesinemisest; nendest tunnustest saab lähtuda ka masinõppel põhinevas klassifitseerimises.

Automaatses klassifitseerimises on kaks peamist suunda (Bird jt 2009). Esimene on juhendatud masinõppimine, mis eeldab, et me teame, millistesse tekstiliikidesse tahame internetist pärinevad tekstid paigutada. Algoritm õpib eelnevalt klassifitseeritud tekstide (treeningmaterjal) peal ning nende teadmiste põhjal loob klassifitseerimismudeli. Teisisõnu, mudel õpib iga tekstiliigi tunnuseid ning peaks olema suuteline uusi seninägemata tekste klassifitseerima. Juhendatud masinõppe meetodeid on erinevaid, nt tõenäosusmudelid, tehisnärvivõrk, tugivektormasinad jm. Teine suund on juhendamata masinõppimine, mis loob "oma tekstiliikide süsteemi" ise, kuna puudub treeningmaterjal, mille pealt õppida. Selle asemel peab mudel andmestikust ise leidma uusi struktuure ja võrgustikke. Juhendamata meetoditena kasutatakse faktor- ja klasteranalüüsi. Käesoleva ülesande – tekstide jaotamine kirjakeele normi järgivaks ja mittejärgivaks – lahendamiseks sobib eelkõige juhendatud masinõpe. Juhendamata masinõpet võiks kasutada tekstiliikide klassifikatsiooni koostamiseks juhul, kui korpuses esindatud tekstiliigid on teadmata.

Klassifitseerimismudelite edukus sõltub suuresti tunnuste valikust ning treeningmaterjali kvaliteedist ja suurusest. Tunnustena võib kasutada sõnehulkasid (ingl *bag of words*), tekstisõnade sõnaliike ja teisi morfanalüüsi tulemusi, süntaktilist infot, tähtede *n*-gramme, lause pikkust, suur- ja väiketähtede suhet jm või nende kombinatsioone (Bird jt 2009). Selles töös kasutatud tunnustest ja klassifitseerimismudelitest on lähemalt juttu osas 4.1.

Enne juhendatud masinõppe rakendamist tuleb aga leida olemasolevate hulgast või luua ise selle materjali ja liigituseesmärkide jaoks sobiv tekstiliikide

[1] http://noisy-text.github.io (27.9.2017).

klassifikatsioon. Selleks on võimalik kasutada "ülevalt alla" või "alt üles" meetodit. "Ülevalt alla" meetodis toetub uurija teoreetilistele printsiipidele või varasematele klassifikatsioonidele (vt Stubbe, Ringsletter 2007, Berniger jt 2008, Laippala jt 2017). "Alt üles" meetodis loovad klassifikatsiooni tavakasutajad, kellel palutakse tekstid liigitada ette antud üldise raamistiku järgi (vt Crowston jt 2011, Egbert, Biber 2013, Egbert jt 2015, Ashegi jt 2016). Selles töös kasutasime testkorpuse loomiseks "ülevalt alla" meetodit, vt osa 4.3.

3. Treening- ja testkorpuse kirjeldus ja töötlus

Järgnevalt tuleb juttu klassifitseerimise katsetes kasutatud korpustest ning nende eeltöötlusest.

3.1. Materjal

Selles töös oleme kasutanud kolme eesti keele korpust: Tasakaalus korpust[2], Koondkorpuse koosseisus olevat Uue meedia korpust[3] ja Estonian Web 2013[4] korpust. Nende korpuste vahel on põhimõtteline erinevus: Tasakaalus korpuse ning Uue meedia korpuse tekstide päritolu ning korpuse tekstiliigiline koostis on täpselt teada, kuid Estonian Web 2013 oma mitte.

Tasakaalus korpus sisaldab ilukirjanduse, ajakirjanduse ja teaduse tekste ning Uue meedia korpus jututubade, foorumite, uudisgruppide ning kommentaaride tekste. Viimasest kasutasime siiski treeningmaterjalina ainult foorumite, uudisgruppide ja kommentaaride tekste, jututoad kui selgelt eripärase keelekasutusega tekstiliigi, mida Estonian Web 2013 korpus ei sisalda, jätsime kõrvale.

Estonian Web 2013 on internetist alla laetud eestikeelsete veebilehtede korpus, mis sisaldab kogu eestikeelset veebi 2013. aasta seisuga ja millest on välja jäetud need tekstid, mis Koondkorpuses[5] juba olemas olid. Sellise korpusekogumise viisi kohta saab lähemalt lugeda nt Miloš Jakubíčeki ja kolleegide (2013) artiklist *The TenTen Corpus Family*. Täisautomaatselt korjatud korpus seab uudse situatsiooni ette: loodud on väga suur korpus, aga pole täpselt teada, millest see koosneb.

Estonian Web 2013 tekstid on poolautomaatselt jagatud seitsme tekstiliigi vahel, millest avaliku halduse (Estonian Web 2013 tekstiliikide tabelis *government*), nn soliidsema ajakirjanduse (*periodical*), informatiivsed (*informative*) ja religiooni (*religion*) tekstid peaksid esindama kirjakeele normi järgivat keelekasutust ning foorumite (*forum*) ja blogide (*blog*) tekstid võiksid oletatavasti sisaldada kirjakeele normi rohkem või vähem järgivaid tekste, st olla sellest seisukohast heterogeensed. Kolmandik Estonian Web 2013 korpusest (arvestades sõnade, mitte dokumentide või domeeninimede arvu) on jäänud liigitamata (tekstiliigi märgendiks on *unknown*). Tekstiliigituse kvaliteet (st kui palju tekste on liigitatud õigesti) on teadmata ja seepärast on ka selles artiklis, võib-olla mõnevõrra lihtsustavalt, väidetud, et Estonian Web 2013 tekstiliigiline koostis on teadmata.

Kasutatud korpustest on küll olemas lemmatiseeritud ja lausestatud versioonid, aga selles artiklis kirjeldatud katsetes kasutasime korpuste lausestatud, kuid

[2] Vt https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/187-grammatikakorpus (26.9.2017).

[3] Vt https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/212-koondkorpus-uus-meedia (26.9.2017).

[4] Vt http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html (26.9.2017).

[5] https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/192-segakorpus (26.9.2017).

morfoloogilise märgenduseta versioone. Estonian Web 2013 märgendamiseks on kasutatud kirjakeele normi järgiva teksti töötlemiseks mõeldud tööriistu, mistõttu märgenduse kvaliteet on kõikuv (aga selle kvaliteedi parandamiseks klassifitseerimist vaja ongi).

Juhendatud masinõppel põhinev klassifitseerija vajab treenimiseks ehk õppimiseks eelnevalt klassifitseeritud korpust. Hea tulemuse saavutamiseks on vaja, et treeningmaterjal oleks võimalikult sarnane selle materjaliga, mille liigitamiseks klassifitseerijat hiljem kasutatakse. Ideaalne oleks kasutada treeningmaterjalina Estonian Web 2013 käsitsi liigitatud allosa. Sellise materjali puudumisel kasutasime treenimiseks tekstikogusid, mis võimalikult lähedaselt esindaksid neid klasse, mille eristamine on töö eesmärgiks: kirjakeele normi järgivad ja mittejärgivad tekstid.

3.2. Materjali eeltöötlus

Kasutatavate korpuste tekste oli tarvis eeltöödelda ja normaliseerida. Programm eemaldas kõigepealt xml-märgenduse ja kustutas kõik reavahetused. Seejärel viis kogu teksti ühele reale ja, mis kõige olulisem, lisas failile kategooria, mis näitab selle kuulumist kirjakeelsesse või mittekirjakeelsesse klassi. Selle tulemusena tekkis igasse faili kolm veergu: kategooria, failinimi ja tekstiline sisu.

Liigse müra vähendamiseks asendasime tekstis veebiaadressid märgendiga <*hyperlink*>, e-posti aadressid märgendiga <*email*> ja igasugused numbrid/arvud märgendiga <*arv*>. Tavapraktikas eemaldatakse üldjuhul ka igasugune punktuatsioon ja/või muudetakse suur algustäht väiketäheliseks, kuid selle ülesande täitmiseks osutaksid need teisendused karuteene. Osalt seetõttu, et internetis (nt kommentaarid, foorumid jm) kiputakse lause alustamist suure algustähega pigem eirama ning kasutatakse kirjavahemärke intensiivsuse markeerimiseks (nt *!!!*, *!????* jm) või jäetakse üldse ära (Hennoste 2013). Seega võiksid suure algustähe ja interpunktuatsiooni kasutamine olla just need tunnused, mis eristavad kirjakeele normi järgivaid ja mittejärgivaid tekste.

4. Eksperimendid

Järgnevalt kirjeldame kasutatud tunnuseid ja klassifitseerimismudeleid ning nende täpsust, mõõdetuna Tasakaalus ja Uue meedia korpusel ning Estonian Web 2013 korpusel.

4.1. Töö käik: tunnuste valik ja klassifitseerimismudelid

Pärast failide eeltöötlust tuli tekstiline sisu muuta masinloetavaks ehk teisendada numbrilisteks tunnusvektoriteks. Tunnustena kasutasime järjestamata sõnade esinemissagedusi ehk sõnehulkasid, mh ka seetõttu, et Sharoff jt (2010) ja Laippala jt (2017) on näidanud, et hoolimata oma lihtsusest on sõnehulkade kasutamine andnud klassifitseerimisel kasutatavate tunnuste võrdluses häid tulemusi. Selle meetodiga luuakse kõikide failide sõnadest sõnastik ning kõikide failide sõnadest ja nende sagedustest vektoresitused. Näiteks olgu korpus, mis koosneb kahest lausest:

(1) Kass joob piima, kuid koer mitte
(2) Valge kass ja must kass nurruvad

Nendest kahest lausest luuakse sõnastik, mis koosneb kümnest sõnast: [*ja*, *joob*, *kass*, *Kass*, *koer*, *kuid*, *mitte*, *must*, *nurruvad*, *piima*, *Valge*, *,* (koma)]. Niisiis on lause (1) ja (2) vektoresitused tabelis 1 järgmised:

Tabel 1. Näitelausete vektoresitused

	ja	*joob*	*kass*	*Kass*	*koer*	*kuid*	*mitte*	*must*	*nurruvad*	*piima*	*Valge*	*,*
(1)	0	1	0	1	1	1	1	0	0	1	0	1
(2)	1	0	2	0	0	0	0	1	1	0	1	0

Tunnuste ekstraheerimiseks ja klassifitseerimismudeli loomiseks kasutasime masinõppe tarkvaraplatvormi *scikit-learn*[6], õppimisalgoritmidest olid esindatud kõik peamised *scikit-learn*'i teegis olevad juhendatud õppe meetodid:

1) multinomiaalne naiivne Bayes (*multinomial Naive Bayes*),
2) logistiline regressioon (*logistic regression*),
3) lineaarne tugivektormasin (*linear support vector*),
4) juhumets (*random forest*),
5) mitmekihiline närvivõrk (*multi-layer perceptron*).

Klassifitseerimismudelite kvaliteeti hinnatakse tavaliselt ristvalideerimisega. See tähendab, et tekstid jagatakse juhuslikult *k* erineva, kuid sama suurusjärguga osa vahel (*k* on kasutaja poolt defineeritud parameeter). Ristvalideerimismeetodis kasutatakse üht osa tekstidest täpselt ühe korra testandmetena ja *k*-1 korda treeningandmetena. Antud ülesandes valisime *k* väärtuseks 10. See tähendab, et õppimisalgoritmi jooksutatakse 10 korda nii, et igas tsüklis kasutatakse 9/10 tekstidest treeningandmete ja 1/10 testandmetena.

Iga tsükkel läbib järgmised sammud:

1) vaadeldavad tekstid (testandmed) eraldatakse ülejäänud tekstidest (treeningandmed);
2) test- ja treeningandmed teisendatakse vektorkujule, kasutades TF-IDF normaliseerimist ja varieerides minimaalset sõne esinemissagedust[7]. Praktikas rakendatakse tunnuste hulgast stoppsõnade[8] eemaldamist, kuid siin kasutasime kaalude skaleerimiseks (*weight downscaling*) TF-IDF[9] (*term frequency–inverse document frequency*) normaliseerimist;
3) treeningandmed liiguvad treenimisetappi ja iga õppimisalgoritmiga luuakse eraldi mudel;
4) iga mudel liigitab testandmete hulka kuuluva teksti ühte võimalikku kategooriasse (kas kirjakeelne või mitte);
5) mõõtmistulemused salvestatakse.

[6] http://scikit-learn.org/stable/ (14.9.2017).

[7] Minimaalne sõne esinemissagedus on parameeter, millega jäetakse kõrvale need sõned, mille tegelik sagedus dokumendis on väiksem kui kasutaja poolt defineeritud esinemissagedus, nt kui min_df = 2, siis sõnastiku loomisel jäetakse kõrvale need sõned, mille sagedus dokumendis on < 2.

[8] Stopp-sõnadeks võib nimetada sõnu, mida soovitakse tekstiandmete hulgast müra vähendamiseks välja filtreerida, nt kõrge esinemissagedusega sõnad.

[9] TF-IDF on statistiline mõõt, mida kasutatakse selleks, et hinnata sõna olulisust korpuses. Sõna olulisus kasvab koos sagedusega, kuid keeles on palju sõnu, millel on väga kõrge esinemissagedus (nt ase-, side- ja kaassõnad), kuid nende olulisus tekstis on üsna väike.

4.2. Treeningkorpus, klassifitseerimismudelite loomine ja nende tulemuste hindamine

Tabelis 2 on esitatud treeningkorpusena kasutatavate tekstide arv ja sõnede maht. Kirjakeele esindajatena kasutasime Tasakaalus korpuse allkorpusi (aja-, ilu- ja teaduskirjandus); kirjakeele normi mittejärgiva keelekasutuse esindajatena Uue meedia korpuse allkorpusi (foorumid, kommentaarid ja uudisgrupid). Klassifitseerimismudelite loomiseks kasutasime lähteandmetena 752 faili, millest 55% esindas kirjakeelt ja 45% kirjakeele normi mittejärgivat keelekasutust. Treeningkorpus koosnes ligikaudu 26,5 miljonist sõnest, k.a punktuatsioon.

Tabel 2. Treeningkorpuse maht

Lähteandmestik	Failide arv		Sõnede arv kokku	Tekstide keskmine pikkus sõnades
Uue meedia korpus	338	197 foorum	~13,6 mln	40 314
		77 kommentaarid		
		64 uudisgrupid		
Koondkorpus	414	138 ilukirjandus	~12,9 mln	31 215
		138 teaduskirjandus		
		138 ajakirjandus		
Kokku	752		~26,5 mln	35 764,5

Järgnevalt esitame ristvalideerimisel saadud erinevate klassifitseerimismudelite täpsused (*accuracy*). Tabelis 3 on esitatud iga mudeli täpsus vastavalt minimaalsele sõne esinemissagedusele.

Tabel 3. Erinevate mudelite täpsused[10]

Mudel	df = 1	df = 2	df = 3	df = 4	df = 5	df = 6	df = 7	df = 8
MNNB	0,940	0,949	0,941	0,943	0,942	0,943	**0,953**	0,952
LR	0,983	**0,988**	0,986	0,984	0,987	0,985	0,983	0,985
LSVC	0,992	0,992	0,991	0,992	0,992	0,992	**0,995**	0,991
RF	0,956	0,962	**0,976**	0,974	0,969	0,965	**0,976**	0,974
MLP	0,996	**0,997**	**0,997**	0,996	0,996	0,996	0,996	0,996

Lühendid: MNNB = multinomiaalne naiivne Bayes, LR = logistiline regressioon, LSVC = lineaarne tugivektormasin, RF = juhumets, MLP = mitmekihiline närvivõrk.

Klassifitseerimismudelite ristvalideerimise eesmärk oli näha, kas ja kuidas õppimisalgoritmid sõnahulkade meetodi ja kasutatava treeningmaterjaliga üldse hakkama saavad. Tabelist on näha, et kõikide klassifitseerimismudelite ennustamise täpsus on küllaltki hea: üle 90% tõenäosusega liigitatakse tekst õigesse klassi. Parima tulemuse andis tehisnärvivõrkudel põhinev algoritm, mis liigitas teksti 99,7% täpsusega õigesse klassi.

[10] Saadav klassifitseerimistäpsus on 10-kordse ristvalideerimise tulemuste aritmeetiline keskmine.

4.3. Testkorpuse loomine

Hindamaks Tasakaalus ja Uue meedia korpuse peal treenitud mudelite kvaliteeti Estonian Web 2013 tekstide klassifitseerimisel, tuli luua eristatava tekstiliigi (kirjakeele normi järgiv *vs*. mittejärgiv tekst) suhtes märgendatud testkorpus. Testkorpusesse valisime tekstid juhuslikult ja esindatud olid kõik Estonian Web 2013 poolautomaatse klassifitseerimise tulemusena eristatud tekstiliigid (vt tabel 4). Kõige rohkem oli testkorpuses perioodikat (34%), seejärel informatiivseid (17%), blogide (15%) ja foorumite (12%) tekste. Avaliku halduse, religiooni ja liigitamata jäänud tekstid olid enam-vähem võrdselt esindatud (8%, 8% ja 6%).

Tabel 4. Testkorpuse jagunemine tekstiliigiti

Estonian Web 2013 tekstiliik	Tekstide arv	% testandmete hulgas
periodicals	74	34
informative	38	17
unknown	14	6
government	17	8
religion	17	8
forum	26	12
blog	34	15
Kokku	220	100

Testkorpuse[11] koostasime "ülevalt alla" meetodil, st lasime üliõpilastel käsitsi liigitada 200 Estonian Web 2013 teksti; iga teksti liigitas kolm märgendajat. Liigitamisel oli võimalik valida kolme vastusevariandi vahel: *kirjakeelne*, *mittekirjakeelne* või *ei oska määrata*. Kirjakeele normi mittejärgivateks tuli liigitada sellised tekstid, milles esines ortograafiavigu, eirati sihipäraselt õigekirjareegleid (väikese algustähe kasutamine lause alguses ja nimede kirjutamisel, kirjavahemärkide mittekasutamine), kasutati lühendeid, esines täishäälikute maksimaalset ärajättu (nt *krt* = *kurat*, *pmst* = *põhimõtteliselt*, *nv* = *nädalavahetus* jm), häälitsusi imiteerivaid sõnu (*hmm*, *mkm*, *aaa*) ja suulisele kõnele omaseid hääldusp̈äraselt kirjutatud võõrlaene (nt *poindile pihta saama*, *tavaar*, *khuul* jm). Raskeks ülesandeks osutus foorumitekstide liigitamine, sest need koosnevad paljude kasutajate postitustest ning kasutajatel on erinev kirjaviis. Probleemsed on ka sellised tekstid, kus kirjakeelsele sisule järgneb kirjakeele normi mittejärgiv sisu (nt samas failis on uudise tekst ja selle kommentaarid). Kirjakeele normi järgivate ja mittejärgivate tekstide vahele on praktikas raske selget joont tõmmata, kuna on raske otsustada, mitu trüki- või ortograafiaviga või kõnekeelsust võib tekstis olla, et seda saaks veel kirjakeelseks liigitada. Neid tekste tulekski pigem vaadelda kui kontiinumit, kus tekstid on rohkemal või vähemal määral (mitte)kirjakeelsed, kuid klassifitseerimismudelite kvaliteedi hindamiseks tuli tekstid kindlakäeliselt kahte klassi liigitada.

See, millise kategooria tekst saab, sõltus enamushääletusest. Teisisõnu, kui tekst sai kolmelt märgendajalt hinnanguks *kirjakeelne*, *kirjakeelne* ja *ei oska määrata*, siis määrati kategooriaks *kirjakeelne*. Juhul, kui inimeste arvamused teksti liigitusest "jäid viiki" (nt tekst sai hinnanguks *kirjakeelne*, *mittekirjakeelne* ja *ei oska määrata*), kaasasime lõpliku otsuse tegemiseks neljanda hinnangu.

[11] Testkorpusesse lisati lisaks 200 käsitsi liigitatud tekstile veel juurde 20 teksti, mis olid varasemalt käsitsi liigitatud.

Tabelis 5 on välja toodud käsitsi liigitamise tulemused: tekstidele määratud tekstiliik (kirjakeelne *vs*. mittekirjakeelne; Estonian Web 2013 olemasolevate tekstiliikide kaupa) ja märgendajate üksmeel selle määramisel.

Tabel 5. Käsitsi liigitatud testkorpus

Tekstiliik Estonian Web 2013 korpuses	Hinnangute kooskõla	Kategooria	Hinnangute kategooriad kokku
blog	19 (65%)	kirjakeelne	9 (31%)
		mittekirjakeelne	20 (69%)
forum	14 (66%)	kirjakeelne	1 (5%)
		mittekirjakeelne	20 (95%)
government	14 (88%)	kirjakeelne	16 (100%)
		mittekirjakeelne	0
informative	24 (69%)	kirjakeelne	16 (46%)
		mittekirjakeelne	19 (54%)
periodical	55 (80%)	kirjakeelne	48 (70%)
		mittekirjakeelne	21 (30%)
religion	11 (69%)	kirjakeelne	16 (100%)
		mittekirjakeelne	0
unknown	9 (64%)	kirjakeelne	5 (36%)
		mittekirjakeelne	9 (64%)
Kokku	146 (66%)		200 (100%)

Tabelist 5 on näha, et tekstid, mille kategooria osas olid märgendajad üksmeelel, moodustasid kõikidest tekstidest 66% (146 teksti), ja tekstid, mille kategooria osas olid märgendajad eri arvamusel, moodustasid kõikidest tekstidest 44% (54 teksti). Kõige suurem märgendajatevaheline üksmeel oli avaliku halduse (88%) ja perioodika (80%) tekstide seas. Ülejäänud tekstiliikide osas jäi märgendajatevaheline üksmeel samasse suurusjärku (64%–69%).

Seega kirjakeele normi järgivad Estonian Web 2013 olemasoleva liigituse järgi avaliku halduse ja religiooni tekstid, samas kui perioodika ja informatiivsete tekstide hulgas on nii kirjakeele normi järgivaid kui ka mittejärgivaid tekste. Perioodika tekstidest liigitati 70% ja informatiivsetest tekstidest vaid 46% kirjakeelseteks. Blogitekstide hulgas oli kirjakeele normi järgivaid tekste 31% ja kirjakeele normi mittejärgivaid tekste 69%. Foorumitekstid liigitati valdavalt kirjakeele normi mittejärgivateks (95%). Estonian Web 2013-s liigitamata jäänud tekstidest 64% järgivad kirjakeele normi ning 36% mitte.

Eelneval on kaks võimalikku seletust: kas perioodika ja informatiivsete tekstide keelekasutus ongi kirjakeele normi järgimise suhtes heterogeenne ja/või on nendesse tekstiliikidesse Estonian Web 2013 poolautomaatse liigitamise käigus paigutatud ka sinna mittekuuluvaid tekste. Siinkohal pole meie ülesanne poolautomaatsel viisil saadud Estonian Web 2013 tekstiliikide ümberklassifitseerimine, vaid eesmärgiks on anda ülevaade, millest loodud testkorpus koosneb.

4.4. Katsed Estonian Web 2013 testkorpusega

Estonian Web 2013 tekstide testkorpuse klassifitseerimisel kasutasime samu õppimisalgoritme, mis olid näidanud head tulemust Tasakaalus korpuse ja Uue meedia korpuse tekstide klassifitseerimisel: multinominaalne naiivne Bayes (MNNB), logistiline regressioon (LR), lineaarne tugivektormasin (LSVC), juhumets (RF) ja mitmekihiline närvivõrk (MLP). Katse tulemused on toodud tabelis 6, mis esitab erinevate klassifitseerimismudelite täpsused käsitsi liigitatud Estonian Web 2013 alamhulgal. Tabeli veergudel on eraldi välja toodud iga mudeli täpsus minimaalsete sõnede esinemissageduste järgi vahemikus 1–10.

Tabel 6. Mudelite täpsused Estonian Web 2013 testkorpuse klassifitseerimisel

Mudel	df = 1	df = 2	df = 3	df = 4	df = 5	df = 6	df = 7	df = 8	df = 9	df = 10
MNNB	0,677	0,686	0,686	0,691	0,695	0,695	0,705	0,709	0,709	**0,714**
LR	**0,645**	0,595	0,6	0,595	0,595	0,605	0,595	0,595	0,605	0,591
LSVC	0,591	**0,65**	0,645	0,645	0,641	0,645	0,645	0,641	0,632	0,623
RF	0,486	0,509	0,523	0,509	0,514	0,527	**0,595**	0,5	0,5	0,491
MLP	0,727	0,727	0,727	0,727	0,732	0,736	0,736	0,736	**0,741**	0,732

Lühendid: MNNB = multinomiaalne naiivne Bayes, LR = logistiline regressioon, LSVC = lineaarne tugivektormasin, RF = juhumets, MLP = mitmekihiline närvivõrk.

Näeme, et võrreldes Tasakaalus ja Uue meedia korpuste tekstide klassifitseerimisega on klassifitseerimismudelite täpsused Estonian Web 2013 testkorpuse peal tunduvalt langenud (vrd tabeleid 3 ja 6). Kõige kehvema tulemuse sai juhumetsa (RF) õppimisalgoritm täpsustega 0,486–0,595 (parim tulemus min_df = 7), mis üldistatult tähendab, et mudel paigutab iga teise teksti valesse klassi. Tugivektormasina (LSVC) ja logistilise regressiooni (LR) mudelid klassifitseerivad tekste enam-vähem sama täpsusega (0,650 ja 0,645) ehk iga kolmas tekst liigitatakse valesse klassi. Parima tulemuse saavutas jällegi tehisnärvivõrkudel põhinev õppimisalgoritm, mille parim täpsus oli 0,741, mis tähendab, et mudel paigutab iga neljanda teksti valesse klassi.

Nende tulemuste põhjal võib oletada, et Tasakaalus korpuse ja Uue meedia korpuse põhjal ei saa hästi modelleerida kogu Estonian Web 2013 tekstilist mitmekesisust. Teiseks võimalikuks tulemuste halvenemise seletuseks on see, et interneti nn kasutaja loodud sisu aastal 2013 (st Estonian Web 2013 korpuses) erineb aastate 2000–2008 keelekasutusest, mida sisaldab Uue meedia korpus. Seda oletust toetab ka Sabiina Haiba bakalaureusetöö (2016), milles järeldatakse, et netikeel esiteks muutub kiiresti ja teiseks muutub pigem kirjakeelsemaks.

Selleks, et mudeli täpsust ja kvaliteeti tõsta, tuleks katsetada uute tunnuste lisamisega (nt lõigu pikkus lausetes ja lause pikkus sõnades mõõdetuna, suur- ja väiketähtede suhe, kirjavahemärkide ja tähtede suhe). Lisaks tuleks mõelda ka treeningandmete hulga suurendamisele. Kui võrrelda klassifitseerimismudelite tulemusi treening- ja testkorpusel, siis kõigi täpsus langes ning see võib olla põhjustatud treeningandmete ühetaolisusest ja sellest, et treeningkorpuse ja testkorpuse tekstid on liialt erinevad. Üheks tulemuse parandamise võimaluseks oleks seega lisada treeningkorpusesse testkorpusele sarnasemaid tekste. Selleks võiks mingi lihtsa meetodiga (nt URL-i järgi) välja võtta Estonian Web 2013 korpusest kõige

prototüüpsemad kirjakeele normi järgivad *vs.* mittejärgivad tekstid ja kasutada neid treeningetapil.

5. Kokkuvõte ja edasiarenduse võimalused

Artiklis kirjeldasime eestikeelsete veebitekstide automaatse liigitamise katseid ning andsime ülevaate erinevate klassifitseerimismudelite kvaliteedist treening- ja käsitsi liigitatud testkorpuse peal. Meie praktilisest vajadusest lähtuv ülesanne oli luua tööriist, mis suudaks liigitada eestikeelseid veebitekste kirjakeele normi järgivasse või mittejärgivasse klassi.

Treeningkorpusena kasutasime Tasakaalus ja Uue meedia korpust ning testkorpusena Estonian Web 2013 käsitsi liigitatud allkorpust. Klassifitseerimiseks kasutasime juhendatud masinõppe algoritme. Nendest algoritmidest töötas sõnehulkade meetodiga kõige paremini mitmekihiline närvivõrk täpsusega 0,741.

Testkorpuse loomine tekstide käsitsi liigitamise teel näitas, et kuna üleminek kirjakeele normi järgivatelt tekstidelt mittejärgivatele on sujuv, siis tuleks ka automaatsel klassifitseerimisel rakendada mitte ranget binaarset klassifitseerimist, vaid lasta klassifitseerimismudelitel prognoosida teksti kuulumist ühte või teise klassi, st anda iga sisendteksti kohta välja ühte *vs.* teise klassi kuulumise tõenäosus.

Viidatud kirjandus

Asheghi, Noushin Rezapour; Sharoff, Serge; Markert, Katja 2016. Crowdsourcing for web genre annotation. – Language Resources and Evaluation, 50 (3), 603–641. https://doi.org/10.1007/s10579-015-9331-6

Berninger, Vera; Kim, Yunhyong; Ross, Seamus 2008. Building a document genre corpus: A profile of the KRYS I corpus. – Proceedings of the BCS-IRSG Workshop on Corpus Profiling, London, UK, October 18.

Biber, Douglas 1988. Variation Across Speech and Writing. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas 1995. Dimensions of Register Variation: A Cross-linguistic Comparison. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Biber, Douglas; Conrad, Susan 2009. Register, Genre, and Style. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Bird, Steven; Klein, Ewan; Loper, Edward 2009. Learning to Classify Text. – Natural Language Processing with Python. http://www.nltk.org/book/ch06.html (10.9.2017).

Crowston, Kevin; Kwaśnik, Barbara; Rubleske, Joseph 2011. Problems in the use-centered development of a taxonomy of web genres. – Genres on the Web. Computational Models and Empirical Studies 42. New York: Springer, 69–84. https://dx.doi.org/10.1007/978-90-481-9178-9_4

Egbert, Jesse; Biber, Douglas 2013. Developing a user-based method of register classification – Proceedings of the 8th Web as Corpus Workshop, WAC-8 2013, 16–23.

Egbert, Jesse; Biber, Douglas; Davies, Mark 2015. Developing a bottom-up, user-based method of web register classification. – Journal of the Association for Information Science and Technology, 66 (9), 1817–1831. https://doi.org/10.1002/asi.23308

Haiba, Sabiina 2016. Kuidas on netikeel muutunud aastatel 2001–2008? Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut. http://hdl.handle.net/10062/56225

Hennoste, Tiit 2000. Eesti keele allkeeled. – T. Hennoste (Toim.), Tartu Ülikooli eesti keele õppetooli toimetised 16. Tartu: Tartu Ülikooli Kirjastus, 9–57.

Hennoste, Tiit 2013. Kuule ma eemale nüüd. – Sirp (46), 40.
Jakubíček, Miloš; Kilgarriff, Adam; Kovář, Vojtěch; Rychlý, Pavel; Suchomel, Vít 2013. The TenTen Corpus Family. – Lancaster, 7th International Corpus Linguistics Conference CL 2013, 125–127.
Kallas, Jelena; Koppel, Kristina; Tuulik, Maria 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel ['New possibilities in corpus lexicography based on the example of the Estonian Collocations Dictionary']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 75–94. https://doi.org/10.5128/ERYa11.05
Kasik, Reet 2007. Sissejuhatus tekstiõpetusse. E. Uuspõld (Toim.). Tartu: Tartu Ülikooli Kirjastus.
Laippala, Veronika; Luotolahti, Juhani; Kyröläinen, Aki-Juhani; Salakoski, Tapio; Ginter, Filip 2017. Creating register sub-corpora for the Finnish Internet Parsebank. – Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa 2017, Gothenburg, Sweden, 152–161.
Santini, Marina 2007. Automatic Identification of Genre in Web Pages. Dissertation. University of Brighton, Computational Linguistics.
Sharoff, Serge; Wu, Zhili; Markert, Katja 2010. The Web Library of Babel: Evaluating genre collections. – Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 17–23.
Stubbe, Andrea; Ringlstetter, Christoph 2007. Recognizing genres. –Abstract Proceedings of the Colloquium "Towards a reference corpus of web genres", Birmingham, UK, July 27.
Särg, Dage 2015. Internetikeele süntaktiline analüüs kitsenduste grammatikaga. Magistritöö. Tartu Ülikool. http://hdl.handle.net/10062/47666

Võrgumaterjalid

Estonian Web 2013. http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html (26.9.2017).
Koondkorpus. https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/192-segakorpus (26.9.2017).
Scikit-learn'i teek. http://scikit-learn.org/stable/ (14.9.2017).
Tasakaalus korpus. https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/187-grammatikakorpus (26.9.2017).
Uue meedia korpus. https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/212-koondkorpus-uus-meedia (26.9.2017).
Workshop on Noisy User-Generated Text. http://noisy-text.github.io (27.9.2017).

Kristiina Vaigu (Tartu Ülikool) uurimisvaldkondadeks on morfoloogia, korpuslingvistika ja arvutilingvistika.
Tartu Ülikool, arvutiteaduse instituut, Liivi 2, 50409 Tartu, Estonia
kristiina.vaik@ut.ee

Kadri Muischneki (Tartu Ülikool) teaduslikud huvialad on korpuslingvistika, eesti keele süntaktiline struktuur ning automaatne süntaktiline analüüs.
Tartu Ülikool, arvutiteaduse instituut, Liivi 2, 50090 Tartu, Estonia
kadri.muischnek@ut.ee

CLASSIFYING ESTONIAN WEB TEXTS

Kristiina Vaik, Kadri Muischnek
University of Tartu

Due to the size of the Internet and the multitude of traditional and new genres there has been an increasing interest in automatic genre classification. Labelling texts in natural language processing is essential because this allows us to select more appropriate language models for the analysis. The aim of the article is to describe and present the results of automatically classifying Estonian Web 2013 texts. We evalued the quality of different classification models on our training and manually labelled test set.

Most of the research on automatic classification has focused on classifying multiple genres, while our objective was to do a binary classification. We set out to classify Estonian Web 2013 texts based on whether they are canonical or not. For training we used the Balanced Corpus to represent canonical language and the New Media Corpus to represent non-canonical language. Due to the non-availability of a binary labelled subcorpus of Estonian Web 2013 texts, we compiled it ourselves by manually labelling it. For classification we used different supervised machine learning algorithms and for features a simple Bag of Words method. The results obtained from the preliminary experiments show that neural networks outperformed other machine learning algorithms achieving over 0.7 on accuracy.

The overall results of this study indicate that in order to increase the accuracy of the classifiers, new features should be added (e.g POS count, sentences per paragraph, words per sentence, uppercase and lowercase letters per sentence etc.). Our best model, the neural network classifier, achieved an accuracy of 0.99 on a training set but only a little over 0.74 on the test set. This suggests that future work requiers a bigger and more appropriate training set. The manually labelling task showed us that the transition from canonical to non-canonical is very smooth. Current models produce a score between 0 and 1, defining if the item belongs to a class or not. Therefore, the classification models must be programmed to be more predictive so that the predictions can be tuned by selecting a threshold.

Keywords: corpus linguistics, automatic classification, natural language processing, machine learning, genre, corpus, Estonian