

TALLINNA REAALKOOL

ADEL MÜÜRSEPP

136.A, REAAL-PROGRAMMEERIMISE ÕPPESUUND

KIRJALIKU KEELEOSKUSE TASEME MÄÄRAMINE SÕNALIIKIDE PÕHJAL

JUHENDAJAD HANNA BRITT SOOTS JA JAAGUP KIPPAR

SISSEJUHATUS

Masinõppe arendamise tulemusena on võimalik keeleõppijate kõne ja kirjalike tekstide keeleoskustaset hinnata automaatselt. Märkimisväärsed edu on saavutatud ingliskeelsete tekstide automaathindamisel, kuid eestikeelsete tekstide taseme määramiseks hetkel reaalselt kasutusel olev lahendus puudub. Kuna eesti keel erineb oluliselt suurema kõnelejaskonnaga indoeuroopa keeltest, ei ole neile keeltele arendatud keeletehnoloogia eesti keelele kohandatav. Seetõttu pole eesti keele automaattöötamiseks kasutatav keeletehnoloogia samaväärselt välja arendatud. Varem on uuritud näiteks käänete tekstis esinemise ja tekstipikkuse seoseid teksti keeleoskustasemega, kuid puudub põhjalikum teadmine sõnaliikide ja sõnatüvede andmete sobivusest keeleoskustaseme ennustamisel masinõppemudelites.

Haridus- ja teadusministeeriumi arengukavad („Eesti elukestva õppe strateegia 2020“, „Eesti keele arengukava 2011–2017“ – pikendatud aastani 2021) tõstavad esile vajaduse luua paindlikke lahendusi eesti keele kui teise keele e-õppeks ja keeleoskuse kontrollimiseks. Eesti keele kestlikkuseks on tähtis arendada muu emakeelega elanikele suunatud keeleõppel kasutatavat keeletehnoloogiat. Selle väljatöötamine on oluline hindajatele, et toetada objektiivset hindamist, kuid eelkõige (ise)õppijatele, näiteks selleks, et valmistuda keeletaset määravate testide jaoks, et teada oma taset ja korduvaid vigu. Edukas automaatne keeleoskustaseme määramine on samuti oluline üldiselt eesti keele jaoks kasutatava keeletehnoloogia arendamiseks.

Siinse töö peamine uurimisküsimus on: Kuidas kõige täpsemini eristada keeleoskustasemeid, kasutades masinõpet ja analüüsides sõnaliikide kasutust?

Uurimistööl on kaks eesmärki. Üks eesmärk on leida, millisel viisil sõnaliikide andmeid esitades on võimalik saavutada täpsem tulemus tekstide taseme hindamisel. Uurimistöös kasutatakse kolmel viisil esitatud andmeid: sõnaliigi osakaal sõnedes ehk tekstisõnades, erineva tüvega sõnade osakaal kõigis tekstisõnades sõnaliigiti ja erineva tüvega sõnade osakaal 300 tekstisõnas sõnaliigiti. Teine eesmärk on leida tekstide klassifitseerimiseks kõige kõrgema ennustamise täpsusega masinõppemudel. Uurimistööl on kaks hüpoteesi. Üks hüpotees on, et keeleoskustaset on võimalik kõige täpsemini ennustada andmetega, kus on leitud erineva tüvega sõnade osakaal 300 tekstisõnast sõnaliigiti. Teine hüpotees on, et kolme andmestiku lõikes kõige kõrgema keskmise ennustamise täpsusega on otsustusmetsa masinõppemudel. Uurimistööl on praktiline väärtus Tallinna Ülikooli eesti vahekeele (õppijakeele) korpuse automaathindamise rakenduse arendamise jaoks. Uurimistöö allikatena on kasutatud Eesti vahekeele korpuse tekstide võrdleval analüüsil põhinevaid uurimusi. Masinõppemudelite ehituse teoreetilise aluse loomiseks kasutati masinõppealaseid teadusartikleid. Masinõppemudelite rakendamise kohta saadi infot tarkvarateekide dokumentatsioonist.

Uurimistöös kasutatav keelekorpus on valim, mis võimaldab teha üldistusi kirjaliku eesti õppijakeele kohta. Töös kasutatud tekstide kogumiseks kasutati dokumentaalvaatlust. Kõrgeima ennustamise täpsusega mudeli ja sobivaima andmestiku leidmiseks kasutati eksperimendi meetodit. Parimad masinõppemudelid leitakse mõõtes iga mudeli täpsust testandmete tekstide keeleoskustaseme määramisel.

Uurimistöö on jaotatud neljaks peatükiks. Esimeses peatükis antakse ülevaade senisest keeleoskustasemete võrdlevast uurimisest eesti keeles. Teine peatükk annab ülevaate masinõppest ja kirjeldab empiirilises osas kasutatavaid masinõppemudeleid. Kolmandas peatükis kirjeldatakse metoodikat ning neljandas analüüsitakse saadud tulemusi.

Autor tänab enda juhendajaid Jaagup Kipparit, kes oli suureks toeks uurimuse tehnilise teostuse osas, ja Hanna Britt Sootsi, kes andis nõu masinõppe rakendamise küsimuste korral ning andis näpunäiteid uurimistöö kirjutamiseks. Tänu väärrib Tallinna Ülikooli Eesti vahekeele korpuse töörühm, kes andis võimaluse panustada oma käimasolevasse teadus- ja arendusprojekti ning juhised korpusesse kogutud tekstide kasutamiseks, ning eriti Kais Allkivi-Metsoja, kelle detailsed soovitusel olid suureks abiks.

SISUKORD

SISSEJUHATUS	1
1. KEELEOSKUSTASEMETE VÖRDLEV UURIMINE EESTI KEELES	5
1.1. Tasemeid eristavad keelelised tunnused.....	5
1.2. Automaathindamise katsetused	6
1.3. Vaadeldavate tunnuste valik	7
2. MASINÕPPEMUDELID	9
2.1. Masinõpe	9
2.2. Töös kasutatavad algoritmid	10
3. METOODIKA.....	16
3.1. Tekstide tasemete määramine ja hoiustamine	16
3.2. Andmete töötlus	18
3.3. Masinõppemudelite treenimine ja mõõtmine	19
4. TULEMUSTE ANALÜÜS.....	22
4.1. Masinõppemudelite analüüs.....	22
4.2. Andmestike analüüs.....	27
KOKKUVÕTE	30
RESÜMEE.....	32
ABSTRACT	33
KASUTATUD MATERJALID	34
Lisa 1. Lingvistiliseks analüüsiks kasutatavad tunnused	39
Lisa 2. Algoritmide klassifitseerimine sarnasuse alusel.....	41
Lisa 3. Andmete ja keeleoskustasemete korrelatsioonimaatriksid.....	44
Lisa 4. Andmebaasi skeemi taastamine	47
Lisa 5. Tekstide andmed	48
Lisa 6- Andmete töötlemise kood	53

Lisa 7. Sõnaliikide tähistused	55
Lisa 8. Masinõppemudelite treenimise kood.....	57
Lisa 9. Katsete eksimismatriksid	58
Kinnitusleht.....	66

1. KEELEOSKUSTASEMETE VÖRDLEV UURIMINE EESTI KEELES

1.1. Tasemeid eristavad keelelised tunnused

Eestikeelses tekstis mõõdetavad lingvistilised tunnused võib jagada kolme rühma: grammatika (morfoloogilised ja süntaktilised tunnused), sõnavara ja sidusus (leksikaalsed, diskursiivsed, semantilised tunnused) ning täpsus (tekstitasandi ja sõnatasandi keelevead) (lisa 1). Peale selle saab vaadelda teksti üldise keerukusega seotud tunnuseid, nagu teksti pikkus sõnades, sõnade ja lausete keskmine pikkus. Eesti keele oskuse taset aitavad üldiselt ennustada eesti keelele iseloomulikud grammatilised tunnused, näiteks käänete arv ja tegusõna vormide mitmekesisus (Vajjala, Lõo, 2013). Sowmya Vajjala ja Kaidi Lõo on leidnud kümme kõige informatiivsemat tunnust keeleoskustaseme ennustamisel (tabel 1).

Tabel 1. Eestikeelsete tekstide 10 kõige informatiivsemat tunnust keeleoskustasemete eristamisel. Kaldkriipsuga on tunnuste nimetustes tähistatud jagamistehet. (Lõo, Vajjala, 2014)

Tunnus	Tunnuse rühm
Unikaalsete sõnade arv / kahekordse tekstisõnade arvu ruutjuur	Sõnavara
Unikaalsete sõnade arv / tekstisõnade arvu ruutjuur	Sõnavara
Ainsuse ja mitmuse 2. pöördes tegusõnad	Grammatika
Alistavad sidesõnad	Grammatika
Isikulises tegumoes tegusõnad	Grammatika
Unikaalsete tegusõnade arvu ruut / tegusõnade arv tekstis	Sõnavara
Eri käänete arv teksti kohta	Grammatika
Unikaalsete tegusõnade arv / kahekordse tegusõnade arvu ruutjuur	Sõnavara
Sidesõnad	Grammatika
Hüüdsõnad	Grammatika

Keeleoskustaseme ennustamisel osutusid informatiivseteks sõnavara näitavad tunnused, mida mõõdeti unikaalsete sõnade osakaaluga. Keeleoskustaseme ennustamisel saavutati kõrge täpsus, kui kasutati andmeid käänete arvu kohta. Käänete arvu kõrge informatiivsus on seletatav eesti keele käänete rohkusega. Madalama keeleoskustasemega keeleõppijad kasutavad kirjutamisel vähem eri käändeid, sest keeleoskuse arenedes vormikasutus mitmekesistub. (Eslon, 2011) Sõnaliikide kasutust kirjeldavad andmed, näiteks sidesõnade ja hüüdsõnade korral, andsid masinõppemudeli treenimisel võrdlemisi kõrge ennustamise täpsuse.

1.2. Automaathindamise katsetused

Kaidi Lõo ning Sowmya Vajjala artiklid on teema poolest sarnased uurimistöös käsitletavale probleemile. Lõo 2012. aasta uurimistöö eesmärk oli määrata, kuivõrd täpse tulemuse annab tekstide taseme järgi klassifitseerimine õppijate sõnavara vahelduvust vaadeldes. Tekste klassifitseeriti kolme üldisema keeleoskustaseme vahel: algelise keelekasutuse tähistamiseks A-tase, iseseisva keelekasutuse tähistamiseks B-tase ja vaba keelekasutuse tähistamiseks kasutatav C-tase. Samuti määrati teksti sõnavararikkus 15 erineva tunnuse abil. Saadud andmed olid kasutusel, et treenida järgmiseid masinõppemudeleid: k-lähim naaber (*k nearest neighbors*), otsustuspuu (*decision tree*) ja neurovõrk (*neural network*). Treenitud mudelite seast andis parima tulemuse neurovõrk, mille täpsus oli 56,3%. (Lõo, 2012)

Vajjala ja Lõo 2013. aasta töös uuriti grammatika ja sõnavara tunnuste ning teksti üldise keerukuse kasutamist keeleoskustasemete A, B ja C hindamisel. Töös uuriti otsustuspuu ja lineaarse regressiooni algoritme ning ansambelõpet. Kombineerides kõiki tunnuseid saavutati keeleoskustaseme ennustamise täpsus 59,33%. Kasutades ainult sõnaliikide osakaale tekstis saavutati täpsus 47,33%. See tulemus ei olnud rahuldav, mistõttu võeti kasutusele kahe klassi vaheline astmeline klassifitseerimine. Astmeliste klassifitseerijate täpsused olid 64,7% ja 66,7%. Töös leiti, et keeleoskustaseme ennustamisel on oluline osa morfoloogilistel tunnustel, näiteks käänevorme ja tegusõna vorme kirjeldavatel tunnustel. (Lõo, Vajjala, 2013)

2014. aastal Vajjala ja Lõo avaldatud uurimuses võeti keeleoskustaseme hindamise aluseks Euroopa keeleõppe raamdokumendi (CEFR) spetsiifilisem kuueastmeline skaala (A1, A2, B1, B2, C1, C2) (Euroopa Nõukogu, 2001), keskendudes neljale peamisele, riiklikel tasemeeksamil hinnatavale keeleoskustasemele A2, B1, B2 ja C1. Töös uuriti 78 tunnuse mõju keeleoskustaseme määramisele. Kasutades tugivektormasina algoritmi SMO (*sequential minimal optimization*), oli täpsus keeleoskustasemete ennustamisel 79%. Lineaarse

regressiooni mudel määras keeleoskustaseme 76% täpsusega. Samuti uuriti, kuidas tunnuste vähendamine mõjutab ennustamise tulemust, ning parimaks tasememääramise täpsuseks saadi 78,3%, võttes aluseks 27 tunnust.

Võrdlemisi kõrge hindamistäpsuseni jõuti ka Janek Kossinski bakalaureusetöös. Kossinski bakalaureusetöö keskendus neurovõrkude treenimisele, kasutades andmeid tekstis esinevate sõnaliikide kombinatsioonide, täpsemalt trigrammide ehk kolmikute (nt nimisõna-määrsõna-tegusõna) kohta. Töös ei kasutatud süsteemselt mudelite hüperparameetrite määramist, mistõttu esimestest mudelitest (täpsusega 28,4%) kõrgema täpsusega mudeliteni (71,8%) jõudmiseni kulus kolm päeva. Töös ei treenitud peale neurovõrkude teisi masinõppemudeleid, millega saab seletada Vajjala ja Lõo saavutatud 79-protsendilisest tulemusest madalamat täpsust. Samuti toodi välja, et parem tulemus saavutatakse võrkotsingus süsteemselt eri hüperparameetreid katsetades. (Kossinski, 2018)

1.3. Vaadeldavate tunnuste valik

Keele omandamise käigus areneb õppija sõnavara märkimisväärselt. Seetõttu võib eeldada, et sõnavara mitmekesisust iseloomustavad tunnused on keeleoskustaseme ennustamisel kõrge informatiivsusega. Seda kinnitavad ka Vajjala ja Lõo uurimistulemused (tabel 1). Üldise leksikaalse varieerumise kõrval on nimetatud tegusõnade leksikaalset varieerumist, kuid muude sõnaliikide lõikes ei ole uurimistöodes sõnavara rikkust vaadeldud. Sõnaliikide hulgas on Vajjala ja Lõo andmetel suurima eristusvõimega hüüd- ja sidesõnad. Puudub ülevaade sellest, kuivõrd täpselt võimaldab eri sõnaliikide sagedus teistest tunnustest eraldi vaadelduna prognoosida teksti taset skaalal A2–C1, ning kuidas muutub tasemelt tasemele sõnavara mitmekesisus sõnaliigiti.

Sõnaliikide osakaal on seotud teksti formaalsuse ja sisutihedusega (Kerge, 2012; Kerge *et al.*, 2007). Võib oletada, et kõrgema keeleoskustasemega õppija kasutab suuremale ametlikkusele ja sisulisele keerukusele viitavaid nimisõnu ja omadussõnu ohtramalt. Samuti kasvab eeldatavasti iga sõnaliigi sõnavaraline mitmekesisus. Suuremaid erinevusi võib oodata avatud sõnaliikide (nimi-, omadus- ja tegusõnade) varieerumises, nagu on esile toonud C1-tasemega eesti keele õppijate ja haritud emakeelekõneleajate kirjutiste võrdlus (Allkivi, 2016). Eri tasemega õppijaid võib eristada ka see, kui palju erinevaid sidesõnu ja teksti ilmetavaid määrsõnu kasutatakse. Töös oli võimalik uurida 17 sõnaliiki, mis on tuvastatavad Tartu Ülikoolis arendatud EstNLTK tarkvarateegi abil (lisa 7).

Selleks, et siduda keelekasutuse grammatiline ja sõnavaraline aspekt, on uurimistöös arvesse võetud nii sõnaliikide sagedus kui ka eri sõnatüvede arv sõnaliigi kohta. Sõnatüvede suhtarvu tekstisõnades kui tekste eristava indikaatori puudus on aga see, et pikemate tekstide korral on unikaalsete sõnatüvede arv teksti kohta väiksem, sest sõnatüved võivad hakata korduma (Meunier, 2013, lk 32). Seetõttu võib selline meetod olla ebatäpne erineva pikkusega tekstide võrdlemiseks ka juhul, kui suhtarvud arvutatakse sõnaliikide lõikes eraldi. Kõrgemate keeleoskustasemete tekstid on uuritavas korpuses keskmiselt pikemad kui madalama tasemega tekstid (A2- ja C1-taseme tekstide keskmine pikkus on vastavalt 86,5 ja 691,5 sõna).

Et vähendada tekstipikkuse mõju sõnatüvede suhtelisele esinemissagedusele, kasutati ennustusmodelite loomisel lisaks andmeid, kus arvutati kirjeldatud näitaja 300 tekstisõna kohta. Töös ei soovitud kasutada standardset näitajat 100 tekstisõna kohta, sest 17 tunnuse võrdlemiseks oli selline tekstisõnade arv ebapiisav. Näitaja arvutati 300 tekstisõna kohta, sest see oli lähedane keskmisele tekstisõnade arvule ja vähendas tekstide kokku liitmist. Kõik sama keeleoskustasemega tekstid liideti kokku ja jagati 300 tekstisõnaga tekstideks, mis võeti kasutusele masinõppemudelite treeningandmetena. Sõnaliikide sageduse ja eri sõnatüvede arvu sõnaliigiti uurimiseks võrreldakse uurimistöös kolme andmestikku: iga sõnaliigi osakaal tekstisõnade hulgas (ei sõltu teksti pikkusest), unikaalsete sõnatüvede osakaal iga sõnaliigi korral, unikaalsete sõnatüvede osakaal sõnaliigiti 300 tekstisõna kohta.

2. MASINÕPPEMUDELID

2.1. Masinõpe

Masinõppes kasutatavate algoritmide ülesanne klassifitseerimisprobleemides on jäljendada funktsiooni, mis klassifitseeriks vektorid ühte defineeritud kategooriatest (Osisanwo *et al.*, 2017, lk 128). Algoritmide ülesanne on matemaatilisi meetodeid kasutades modelleerida päriselulist probleemi. Seatud probleemi kohta info saamiseks kasutatakse erinevaid sisendi võtmise viise. Mudeli ehk matemaatilise hüpoteesi koostamiseks on sisendandmete kasutamisest lähtuvalt kaks viisi: juhendamata ja juhendatud masinõpe. Juhendatud masinõppes nimetatakse sisendandmeid treeningandmeteks. Treeningandmetel on kindlaks määratud klass. Juhendatud masinõppe mudel teeb ennustuse treeningandmete andmepunkti klassi kohta. Seejärel võrreldakse ennustatud ja tegelikku klassi ning saadud tagasiside alusel parendatakse matemaatilist hüpoteesi, et saavutada kõrgem mudeli ennustamise täpsus. Treeningprotsess kestab, kuni soovitud ennustustäpsus on saavutatud. (Brainlee, 2019) Sel viisil treenitavad algoritmid sobivad selliste probleemide lahendamiseks, kus treeningandmetel on määratud klassid või kui probleem on kergelt defineeritav ja andmed on võimalik jagada klassidesse. Selliste probleemide näiteks saab tuua rämpsposti tuvastamise ja finantsanalüüsi. Juhendamata masinõppe algoritmide sisendandmetel ei ole määratud klasse. Mudel koostatakse andmestikus esinevaid struktuure analüüsides, eesmärgiga üldistada andmestikku informatiivsete järelduste saamiseks. Juhendamata masinõppe on kasutusel probleemide korral, kus andmete klass on keeruliselt määratav ja treeningandmete kogus suur. Juhendamata masinõppe probleemide näiteks võib tuua pildituvastuse ja muusikastiili määramise.

Uurimistöös kasutatakse peamiselt juhendatud masinõppe alla kuuluvaid algoritme. Juhendatud masinõppe algoritmid sobivad selgelt defineeritud keeleoskustasemete ennustamiseks ning on ka vähem kui 1000 uuritava teksti korral väga täpsed. Juhendamata masinõppe algoritmide seast vaadeldakse kahte neurovõrkude algoritmi: pertseptron ja stohhastiline gradientlaskumine (*stochastic gradient descent*). Järgnevas masinõppealase kirjanduse ülevaates vaadeldakse ainult uurimistöös kasutatavaid masinõppe algoritme ja põhjendatakse, miks enim rakendatavamate algoritmide seast (lisa 2) valiti keeleoskustaseme määramiseks kirjeldatud algoritmid.

2.2. Töös kasutatavad algoritmid

Töö empiirilises osas kasutatavad algoritmid on jagatud sarnasuse alusel rühmadesse (lisa 2). Alapeatükis antakse ülevaade iga rühma algoritmide tööpõhimõttest (tabel 2) ning põhjused, miks algoritm sobib keeleoskustasemete ennustamiseks. Uurimistöö empiirilises osas ei uurita klasterduse algoritme, sest klasterduse algoritmid kasutavad juhendamata masinõpet, et leida iseseisvalt andmestikus mustreid ning sarnanevad juba uuritavate klassifitseerijatega. Kuna algoritmid ei kasuta õppimiseks treeningandmete keeleoskustasemeid, on ebatõenäoline, et saavutatakse kõrge ennustamise täpsus. Empiirilises osas ei kasutata seostepõhiseid algoritme, kuna nende kasutusvaldkond on väga kitsas ja neid ei ole tulemuslikult rakendatud tekstide klassifitseerimiseks. Uurimistöö empiirilises osas ei vaadelda regularisatsiooni rühma kuuluvaid algoritme, sest need sarnanevad tööpõhimõttelt regressiooni rühma algoritmidele ja nende rakendamine tekstianalüüsis pole levinud.

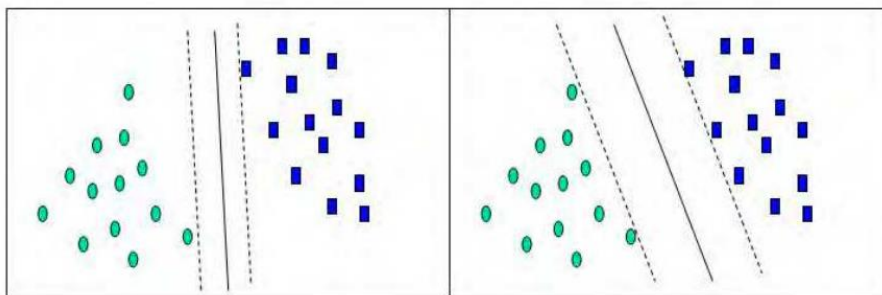
Tabel 2. Empiirilises osas kasutatavad algoritmid sarnasuse alusel rühmadesse jagatuna (Brainlee, 2019)

Algoritmide rühm	Algoritmid
Regressioon	Lineaarne regressioon, kantregressioon
Andmepunktipõhised	Lineaarne tugivektorklassifitseerija, tugivektorklassifitseerija, lähima naabri algoritm, lähima tsentroidi algoritm
Otsustuspuud	Otsustuspuu algoritm
Bayesi algoritmid	Bernoulli jaotusega naiivse Bayesi algoritm, komplementaarse jaotusega naiivne Bayesi algoritm ja multinoomjaotusega naiivne Bayesi algoritm
Neurovõrgud	Stohhastiline gradientlaskumine, pertseptron
Mõõtmete vähendamine	Lineaarne diskriminantanalüüs, ruutfunktsiooni diskriminantanalüüs
Ansambelõpe	Otsustusmets, gradiendi võimendamine, Adaboosti algoritm

Regressioonil põhinevad klassifitseerijad jagavad vektorkujul esitatud tekstid tunnuste lineaarse kombinatsiooni alusel kategooriatesse (Verlinde, *et al.*, lk 1998). Vektorkujul teksti $\vec{x}$ n tunnust on esitatud n -mõõtmelises ruumis. (Taiwo, 2010, lk 24). Lineaarse regressiooni tööpõhimõttest annab ülevaate valem $y = \vec{w} * \vec{x} = \sum_{i=1}^n w_i x_i$, kus leitakse iga tunnuse x_i ja

vastava osakaalu w_i korrutiste summa ning $\vec{w}$ tähistab algoritmi käigus õpitud osakaalude vektorit. Regressioonil põhinevad klassifitseerijad moodustavad kategooriate vahel mitmemõõtmelise otsustuspiiri ehk hüpertasandi. (Taiwo, 2010, lk 25) Lineaarne regressioon on kõige algelisem masinõppemudel ning sobib teiste masinõppemudelite ennustamise täpsuste võrdlemiseks. Töös uuritakse ka kantregressiooni, milles kasutatakse Ridge'i kulufunktsiooni, et vähendada suure osakaaluga mõõtmete tundlikkust väärtuste suhtes.

Juhul, kui andmepunkte on võimalik lineaarselt klassifitseerida, võib eksisteerida mitmeid võrdselt hästi klasse eraldavaid hüpertasandeid. Kui suurema koguse andmete korral valida hüpertasand juhuslikult võimalike hüpertasandite seast, võib mudeli ennustamise täpsus väheneda. Intuiitselt on võimalik mõista, et parim tulemus saavutatakse, kui klassidevaheline ala ehk marginaal on võimalikult lai (joonis 1). (Louppe, 2014, lk 22) Osakaalude vektor $\vec{w}$ on tugivektorklassifitseerija korral võimalik leida valemiga $\vec{w} = \sum_{i=1}^n \alpha_i y_i x_i$, kus $0 \leq \alpha_i \leq C$ ja C on hüperparameeter, mille abil on võimalik määrata, mil määral on marginaalil lubatud andmepunkte katta. (Boser *et al.*, 1992; Cortes, Vapnik, 1995) See on vajalik juhul, kui marginaali sees asuvad eri klasside andmepunktid (Louppe, 2014, lk 22). Uurimistöös kasutatakse lineaarse tugivektorklassifitseerija algoritmi näitena LinearSVC algoritmi. Mittelineaarse hüpertasandi leidva tugivektorklassifitseerija algoritmina kasutatakse SVC algoritmi. Mittelineaarse tugivektorklassifitseerija kernelfunktsioon leitakse võrkotsingu käigus polünoom-, sigmoid- ja radiaalbaasfunktsiooni seast.



Joonis 1. Suure ja väikese marginaalide mudelite võrdlus. Erinevate kategooriate andmepunktid on tähistatud eri värvidega. (Taiwo, 2010)

Lähima naabri algoritmide korral luuakse matemaatiline hüpotees, arvestades andmepunktide paigutust üksteise suhtes. Sellesse kategooriasse kuuluvad algoritmid leiavad kindla arvu treeningandmete andmepunkte, mis on uuritavale testandmete andmepunktile kõige lähemal, ja klassifitseerivad testandmete andmepunkti selle baasil. Arvesse võetavate andmepunktide arv võib olla kasutaja määratud või põhineda vaadeldavat andmepunkti ümbritsevate andmepunktide esinemissagedusel (raadiusepõhine masinõpe). (Scikit-learn'i kodulehekülg)

Uurimistöös kasutatakse määratud andmepunktide arvuga k -lähima naabri ja raadiusepõhist lähima tsentroidi (*nearest centroid*) algoritmi, kauguse leidmiseks kasutatakse eukleidilist ja Manhattani kaugust.

Otsustuspuu algoritmi aluse on üks sõlm, mis jaguneb lahknevateks harudeks. Harud tipnevad lehtede ehk algoritmi määratud klassidega. Iga sõlme juures jagatakse andmepunktid tunnuse väärtuse alusel kaheks. Teksti sõnaliikide andmed on 17-mõõtmelised (igale uuritavale tunnusele lähenetakse kui eraldi mõõtmele), mis tähendab et paljude erinevate tunnuste vaatlemise järjekordade tõttu on võimalike treenitavate andmepuude arv väga suur. Kuna kõige informatiivsema otsustuspuu leidmiseks algoritmi ei leidu, lähenetakse kõige täpsema otsustuspuu leidmisele ahne algoritmi strateegiaga. Igale tunnusele määratakse väärtus, et kirjeldada tunnuse informatiivsust. Alguses võetakse kõige informatiivsem tunnus, seejärel informatiivsusest järgmine tunnus ja jätkatakse samal viisil. Ahne algoritmi täiendav eesmärk parima otsustuspuu leidmisel on otsustuspuu võimalikult väike sügavus, et hoiduda liigsest jaotamisest. (Hyafil, Rivest, 1976) Sügavus on määratud vastava hüperparameetriga.

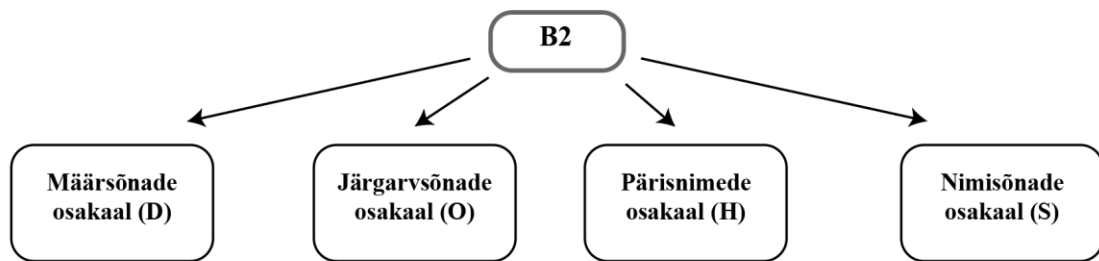
Kui otsustuspuude kihtide arvu mitte piirata, on probleemiks ülesobitus, sest otsustuspuu kihtide arv saab kasvada, kuni iga tunnuse põhjal jagades on viimases kihis üksikult klassifitseeritud andmepunktid. Otsustuspuu kihtide piiramine võib vähendada mudeli täpsust andmete klassifitseerimisel. (Koehrsen, 2018) Selleks, et hoiduda kihtide piiramisest, kasutatakse otsustusmetsa algoritmi. Otsustusmets on mitmetest otsustuspuudest koosnev mudel, mis kuulub ansambelõppe algoritmide hulka. Kaks olulist aspekti otsustusmetsa mudelite loomisel on juhuslikult valitud andmed ja juhuslikult valitud tunnused. Iga otsustuspuu treenitakse juhuslikult valitud andmete põhjal, kusjuures teatud andmed korduvad, mis jätab iga otsustuspuu puhul kasutamata teatud arvu andmepunkte. Kuigi iga individuaalne otsustuspuu ei ole piisava täpsusega klassifitseerimisel, on nende otsustuspuude keskmine tulemus kõrge täpsusega. Otsustusmetsa algoritm on vähem tundlik treeningandmete tekstides sisalduva kõrvalise info suhtes, mis vähendavad mudeli täpsust testandmete tekstide klassifitseerimisel. (Koehrsen, 2018) Uurimistöö andmete kasutamisel valitakse neli tunnust ja iga otsustuspuu treenitakse vaid valitud nelja tunnuse baasil. Otsustusmetsa katses treenitakse sadu tuhandeid otsustuspuuid, mille kombineerimisel saadakse täpsem tulemus (Scikit-learn'i kodulehekülg).

Naiivsed Bayesi algoritmid põhinevad Bayesi valemil. Naiivsete Bayesi algoritmide tööpõhimõte on tõenäosuse põhjal andmete rühmadesse jagamine (Smola, 2008, lk 17). Oluline valem, millest lähtutakse naiivsete Bayesi masinõppemudelite treenimisel on

$$L(x) := \frac{p(B1|x)}{p(B2|x)} = \frac{p(x|B1)p(x)}{p(x|B2)p(x)},$$

kus $L(x)$ on suurem väärtusest 1, kui teksti x B1 taseme tõenäosus on suurem, $L(x)$ on väiksem väärtusest 1, kui teksti x B2-taseme tõenäosus on suurem.

Kuna igal tekstil on 17 tunnust, tuleb arvutada keeleoskustaseme tõenäosus iga tunnuse kohta eraldi. Selleks, et selle arvutamist lihtsustada, tehakse naiivne eeldus (Smola, 2008, lk 22), et tunnuste esinemiste tõenäosused pole üksteise esinemistest mõjutatud. See eeldus jätab tähelepanuta näiteks selle, et tegusõnade suurem esinemissagedus tõstab tõenäoliselt ka sidesõnade esinemissagedust tekstis (joonis 2). Naiivne Bayesi algoritm eeldab näiteks, et B2-keeleoskustaseme ennustamisel puudub seos määrsõnade ja järgarvsõnade osakaalu vahel (joonis 2), kuigi tunnuste esinemise korrelatsioon on peaaegu täielik (lisa 3).



Joonis 2. Tekstis esinevad tunnused ei ole mõjutatud teiste tunnuste esinemise tõenäosusest

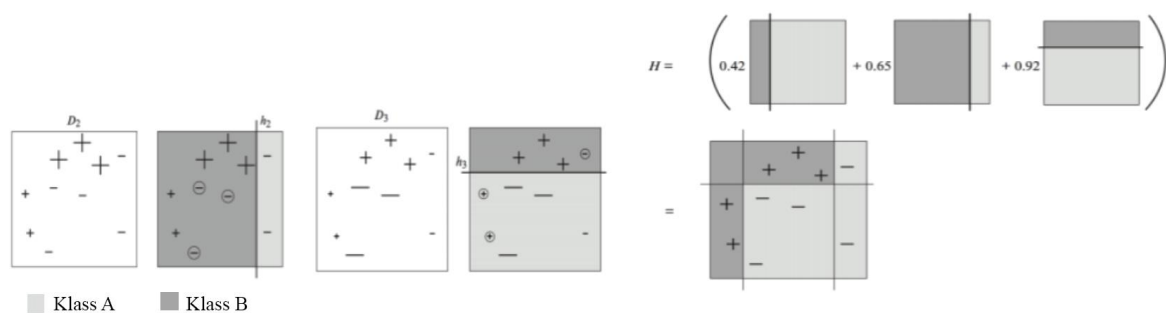
Iga eraldi tunnuse tõenäosuse saab leida, kui vaadelda tunnuse esinemist keeleoskustaseme kohta. Sellisel põhimõttel töötavad valemid on eriti tõhusad tekstide klassifitseerimisel, näiteks kasutatakse selliseid algoritme e-posti rämpsikirjade tuvastamiseks (Smola, 2008, lk 23). Empiirilises osas kasutatakse kolme eri jaotusega algoritmi: Bernoulli jaotusega naiivse Bayesi, komplementaarse jaotusega naiivse Bayesi ja multinoomijaotusega naiivse Bayesi algoritmi.

Neurovõrkude algoritmide seast uuritakse stohhastilise gradientlaskumise algoritmi ja pertseptroni algoritmi. Pertseptron jäljendab neuroni tööpõhimõtet. Treenimise käigus määratakse igale tunnusele 17 tunnusest osakaal. Kui tunnuste osakaalude summa ületab teatud piiri, määrab algoritm vastava klassi. (Raschka 2015) Treenimise käigus kasutatakse 0-1 kaofunktsiooni, st määratakse treeningandmete keeleoskustaseme valesti ennustanud osakaalude kombinatsioonidele karistuseks 1 ja õigesti ennustanud osakaalude kombinatsioonidele karistuseks 0. Pertseptroni algoritmi puuduseks on lineaarse otsustuspiiri

määramine. (Seeger, 2012, lk 18–20) Stohhastiline gradientlaskumine kasutab kärbitud lineaarset kulufunktsiooni ja leiab funktsiooni globaalse miinimumi. Globaalse miinimumi abil leitakse optimaalne osakaalude kombinatsioon. (Seeger, 2012, lk 25–27) Ekstreemväärtuse leidmise kiirust ja täpsust reguleeritakse õpisammu hüperparameetriga. Töös uuritakse ühe kihiga neurovõrke, et leida, kas 958 korpuse teksti kasutamine on piisavalt suur kogus andmeid, et ka mitmekihilisemaid neurovõrke edukalt rakendada.

Lineaarne diskriminantanalüüs (LDA), ruutfunktsiooni kasutav diskriminantanalüüs (QDA) ja peakomponentide analüüs (PCA) on lineaarse transformatsiooni algoritmid. PCA on juhendamata, st see ignoreerib treeningandmete andmepunktide klasse. Vaadeldes andmete jaotumist, arvutab PCA uued mõõtmed, mis suurendavad andmepunktide varieeruvust. See tähendab, et vähem informatiivsetel tunnustel on väiksem mõju algoritmi treenimisel, sest nende osakaalu vähendatakse fiktiivse mõõtme koostamisel. (Raschka 2014) LDA kuulub juhitud masinõppe algoritmide alla, kus arvestatakse andmepunktide klasse. Algoritm arvutab maksimaalselt andmepunkte eristavad mõõtmed ehk lineaarsed diskriminandid. QDA algoritm sarnaneb LDA tööpõhimõttele, kuid sobitab otsustuspiiri ruutfunktsioonile vastavalt. LDA ja QDA algoritmide treenimisel kombineeritakse algoritmid PCA-ga, et võimalikult efektiivselt mõõtmeid vähendada.

Lisaks otsustusmetsa algoritmile uuritakse ansambelõppe rühma algoritmide seast gradiendi võimendamise (*gradient boosting*) ja Adaboosti algoritme. Mõlemad algoritmid kuuluvad võimendamise (*boosting*) algoritmide hulka. Võimendamise eesmärk on iteratiivselt rakendada nõrgema astme klassifitseerijaid uuendatud osakaaludega andmetele ja seejärel järjestada need tulemuslikkust võrreldes enamushääle abil. (Friedman *et al.*, 2000) Lisaks igale klassifitseerijale määratud kaalule määratakse kaal igale andmepunktile, seejuures antakse suuremad kaalud enne korduste alustamist baasklassifitseerija valesti määratud andmepunktile. Kuna rakendatavad otsustuspuud on väga väikese sügavusega, on võimendamise mudelid resistentsed ülesobituvusele ja nende ennustamise täpsus ei vähene suurema korduste arvuga. Enne lõplikusse mudelisse lisamist treenib Adaboosti algoritm alusklassifitseerijad eraldi (joonis 3). Esimese alusklassifitseerija valesti ennustatud andmepunktile määratakse suuremad osakaalud. Järgmise alusklassifitseerija loomisel püütakse õigesti määrata varem valesti määratud andmepunktid, mis on joonisel tähistatud suurema märgiga. Selline tegevus kordub hüperparameetriga määratud korduste arvu kordi. Alusklassifitseerijatele antakse nende täpsuse põhjal osakaal ja need liidetakse lineaarselt kokku, et luua lõplik ennustusmudel.

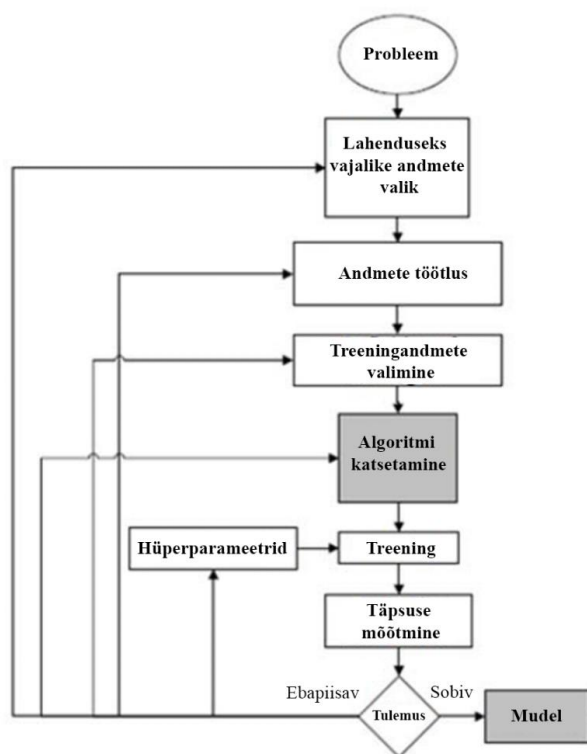


Joonis 3. Vasakul diagrammil eraldi otsustuspuu klassifitseerijate treenimine. Treenimist alustatakse vasakpoolsest otsustuspuust. Paremäl joonisel on näidatud eraldi treenitud klassifitseerijate lisamine lõplikusse mudelisse. (Schapire, Freund, 2012)

Kui AdaBoost arvestab uute alusklassifitseerijate loomisel enim valesti määratud andmepunkte, siis gradiendi võimendamise algoritm arvestab alusklassifitseerijate loomisel enim gradiente. Sarnaselt AdaBoostiga kombineerib gradiendi võimendamise algoritm nõrgad alusklassifitseerijad väljendatuna kujul $F(x) = \sum_{i=1}^n c_i f_i(x) + const$, kus c_i on alusklassifitseerija $f_i(x)$, n on tekstide arv ja $F(x)$ on lõplik ennustusmudel. (Friedman, 2001) Gradiendi võimendamise algoritmi eelis on lisatavate funktsioonide tõttu mittelineaarse hüpertasandi leidmine. Kuna töö maht on piiratud käsitlemaks kõigi potentsiaalsete algoritmide positiivseid ja negatiivseid külgi, on antud peatükk kokkuvõtlik.

3. METOODIKA

Katsete tegemiseks kasutatakse standardset juhendatud masinõppemudeli loomise töökäiku (Kotsiantis, 2007). Iga katse kujutas endast iteratiivset protsessi (joonis 4) ning järgnevalt on kirjeldatud praktilise osa töökäigu iga etappi.



Joonis 4. Juhendatud masinõppemudeli loomise töökäik (Kotsiantis, 2007)

3.1. Tekstide tasemete määramine ja hoiustamine

Uurimistöös kasutati Tallinna Ülikooli eesti vahekeele korpuse (EVKK) tekste. EVKK on eesti keele kui riigikeele (teise keele) ja võõrkeele õppijate kirjalike tekstide elektroonne kogu. (Eslon, 2014) EVKK kombineerib erinevaid alamkorpuse, metateavet tekstide autorite kohta, infot vealiikidest ja veebirakendusi tekstide lingvistiliseks analüüsiks ning võimaldab seetõttu korpust kasutada empiirilises ja rakenduslikku laadi uurimistöös. (Eesti vahekeele korpus) Uurimistöö tekstivalimi moodustavad neli alamkorpust. Esimese, mahukaima alamkorpuse ehk tuumkorpuse tekstide kirjutajad on valdavalt vene emakeelega eesti keele õppijad. Nende kirjalikud tööd on kogutud kahest suurima vene keelt emakeelena kõneleva populatsiooniga Eesti regioonist: Ida-Virumaalt (Narva, Kohtla-Järve, Jõhvi,

Sillamäe) ja Harjumaalt (Tallinn). Lisaks sisaldab tuumkorpus soome, inglise, saksa, ungari, leedu jm emakeelega õppijate töid, kelle jaoks eesti keel on võõrkeel. Teine õppijakeele alamkorpus koosneb riiklikest eksamitöödest, mis on talletatud Innove arhiivis. Digiteeritud on riigieksamite, eesti keele taseme- ja kodakondsuse eksami kirjalikud tööd, põhikooli ja gümnaasiumi eksamitööd. Kolmas õppijakeele alamkorpus sisaldab iga-aastase üleriigilise eesti keele kui teise keele olümpiaadi töid. Neljandas õppijakeele alamkorpuses on akadeemilise õppijakeele tekstid üliõpilaste lühematest või pikematest teadustöödest. (Eslon, 2014)

Tuumkorpuse tekstide esialgsel, kolmeastmelisel keeleoskustaseme määramisel pole kasutatud kvalifitseeritud hindajaid: aluseks võeti õpetajalt saadud metainfo või tekstisisestaja hinnang. Seetõttu võeti ette tuumkorpuse tekstide ja olümpiaaditööde ümberhindamine kuueastmelises süsteemis, mille viisid läbi kolm kvalifitseeritud eksperti (tavapraktikas lubatud miinimum). Ümbervaatamisele ei läinud Innove arhiivi alg-, kesk- ja kõrgtaseme tekstid ning kuueastmelise süsteemi alusel hinnatud eksamitekstid, kuna nende taseme on juba määranud kõrgelt kvalifitseeritud hindajad. Kokkuvõttes sisaldab EVKK 987 ekspertide hinnatud teksti (tabel 3).

Tabel 3. Kuueastmelises süsteemis (A1, A2, B1, B2, C1, C2) on ekspertide hinnang antud kokku 987 tekstile

Keeleoskustase	Sõnade arv	Tekstide arv
A1	76	1
A2	48 742	217
B1	119 830	411
B2	104 183	228
C1	111 435	129
C2	1450	1

Uurimistöös lähtutakse 985 tekstist ainult neist tekstidest, mille tase A2, B1, B2 või C2. Nelja keeleoskustaset uuriti varem tehtud uurimistöödega võrreldavuse huvides. Igal tekstil on korpuses kolm esitusviisi: Wordi dokument, milles on salvestatud tekst koos pealkirja ja tööjuhise, *unicode*-vormingus salvestatud txt-fail, kus pealkirjad ja tööjuhised puuduvad, ning morfosüntaktiliselt analüüsitud tekst. Uurimistöös kasutati tekstide analüüsimisel txt-failivormingut.

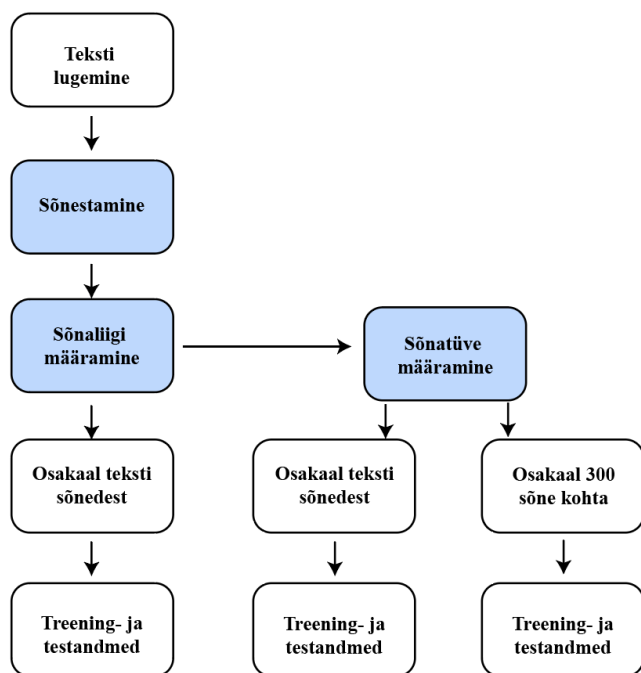
3.2. Andmete töötlus

EVKK tekstid on hoiustatud MariaDB andmebaasis. Uurimuses kasutati psql keelt, mis baseerub SQL päringukeelel. Tekstide ja metaandmete salvestamiseks loodi eraldi Postgres andmebaas. Kuna autorile polnud teada EVKK andmebaasi skeem ehk kuidas on erinevad andmetabelid relatsiooniliselt seotud, koostati andmete veergude pealkirjade järgi andmebaasi oletatav skeem käsitsi. Selleks uuriti, millistele tabelitele vastavad iga tabeli võõrvõtmed. Andmebaasi skeemi loomise eeldus oli võõrvõtmete viidatud andmetabelite olemasolu andmebaasis. Selleks leiti koostatud skeemil andmetabelid, mis ei omanud võõrvõtmeid, ja lisati need andmetabelid DBeaver keskkonnas esimesena (lisa 4). Seejärel kopeeriti psql keelt kasutades EVKK andmed tabelitesse. Loodud andmebaasist laeti iga tekst alla txt-faili kujul, kus failinimi vastas tabelites tekste eristatavatele koodidele. Lisas (lisa 5) on toodud näitena *unicode*-vormingus txt-failina vene keelt emakeelena kõneleva õpilase kirjutatud essee.

Treening- ja testandmestiku jaoks oli vaja kahte sorti töödeldud andmeid: andmetabel, kus on märgitud igale tekstile antud tasemehinnang ja iga sõnaliigi osakaal teksti sõnedest, ning andmetabel, kus on märgitud tekstidele vastav keeleoskustase ja iga sõnaliigi alla kuuluvate unikaalsete sõnatüvede osakaal kõigist selle sõnaliigi tekstisõnadest. Esimesel viisil töödeldud andmete saamiseks oli vajalik teada metaandmete tabeli „dokmeta“ refereeritavas tabelis (lisa 5) sisalduvat tasemehinnangut ning tekstide iga üksikut sõna ehk unigrammi iseloomustava andmetabeli „ngram1“ sõnaliigiinfot (lisa 5). Andmetabelite „dokmeta“ ja „ngram1“ liitmise kood on toodud lisas (lisa 6). Sel viisil saadud andmetabelis oli info kõigi keeleoskustasemetega tekstide kohta. Kuna uurimistöös võrreldi ainult A2-, B1-, B2- ja C1-tasemega tekste, filtreeriti vastavate keeleoskustasemetega tekstid andmetabelist välja (lisa 6). Seejärel salvestati saadud andmetabel csv-vormingus.

Selleks, et leida sõnaliigiti sõnatüvede ja kõigi tekstisõnade suhtarv, töödeldi valimisse kuuluvat 985 teksti Pythoni EstNLTK teegi abil (joonis 5). Kasutades pandas tarkvarateeki, loodi igast csv-vormingus tekstist EstNLTK Text objekt. Text objekti sisu jagati eraldi tekstisõnadeks ehk sõnestati. Selles etapis oli märgata, et EstNLTK tarkvarateek ei tuvastanud kõiki sõnesid õigesti, kuna madalamate keeleoskustasemetega tekstidel puudusid tühikud (lisa 5). Sõned, mis EstNLTK abil valesti tuvastati, määrati pärisnimedeks, mistõttu võib pärisnimede osakaal osutada tegelikust suuremaks. Seejärel määrati tekstis iga tekstisõna sõnatüvi ja leiti sõnaliik (lisa 6). EstNLTK teegi kasutamisel esines ka probleem, et EstNLTK määras mõne sõna jaoks kaks erinevat sõnaliiki, kuna kirjutatud sõne sõnaliik ei olnud üheselt määratav (nt saab sõna palgata esineda nimisõna ja tegusõnana). Kuna iga sõna, millele määrati kaks sõnaliiki, läbivaatamine ja käsitsi määramine ei olnud mõistlik, otsustati valida

esimesena määratud sõnaliik. Esimese määratud sõnaliigi valimine ei mõjuta eeldatavalt andmestiku usaldusväärsust väga oluliselt, kuna selliseid juhte oli andmestikus vähe, arvestuslikult alla ühe protsendi. Sõnatüvede loendamiseks eemaldati EstNLTK tarkvarateeki kasutades iga sõnaliigi sõnedest kõik korduvad tüved, et iga sõnatüvi esineks iga teksti sõnaliigi tekstisõnade seas vaid ühe korra (lisa 6).



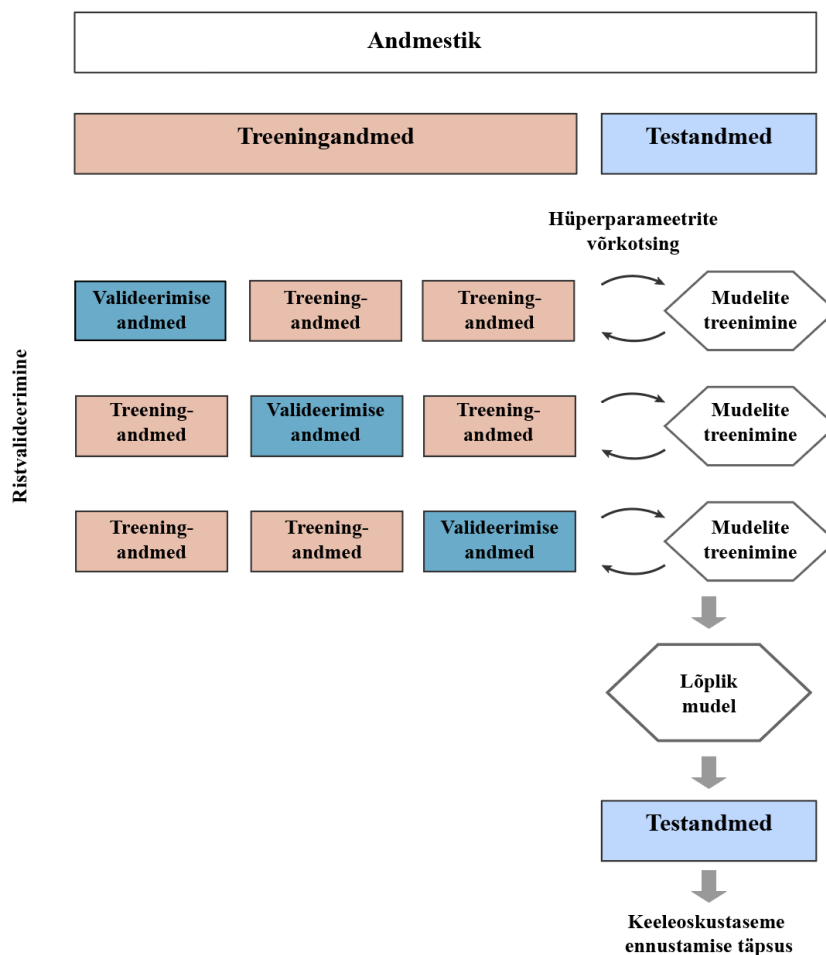
Joonis 5. Andmete töötlus. Sinisega on tähistatud etapid, kus kasutati EstNLTK tarkvarateeki.

Peale tekstide EstNLTK tarkvarateegi abil töötlemist arvutati välja kolme andmestiku jaoks vajalikud sõnaliigi sõnede või sõnaliigi eri sõnatüvede osakaalud. Masinõppemudelite treenimiseks kasutatavasse andmetabelisse lisati arvutatud andmetele ka iga teksti keeleoskustase. Selleks, et lisada iga teksti kohta keeleoskustase, kasutati varem töö käigus loodud andmetabelit, kus oli vajalik teave olemas. Iga teksti keeleoskustase ja sõnaliikide osakaalud salvestati pandase DataFrame-tüüpi andmetabeli objektina (lisa 5). Selleks, et masinõppemudelil oleks võimalik teksti keeleoskustaseme ennustamise funktsiooni modelleerida, kodeeriti keeleoskustasemed numbrilisteks väärtusteks (lisa 6). Saadud andmed salvestati csv-vormingus (lisa 5). Andmed jagati treening- ja testandmeteks.

3.3. Masinõppemudelite treenimine ja mõõtmine

Masinõppemudeli treenimisel on optimaalne testandmete osakaal andmetest 30% (Korjus et al., 2016, lk 11). Treeningandmed moodustasid andmestikust 70% ning need olid kasutusel mudeli treenimiseks. Testandmed moodustasid andmestikust 30% ja nende põhjal leiti

treenitud mudelite keeleoskustasemete ennustamise täpsus. Treeningandmed olid valitud juhuslikus järjestuses. Selleks, et võimaldada masinõppemudelite treenimist erinevate andmete põhjal ja vähendada ülesobituvust, kasutati ristvalideerimist (joonis 6). Kõik treeningandmed jagati omakorda juhuslikult kolme võrdsesse rühma. Seejärel treeniti mudeleid kolme erineva treeningandmete rühmaga. Igal treeningkorral kasutati 2/3 treeningandmetest treenimiseks ja üle jäänud 1/3 andmete ehk valideerimise andmete abil arvutati hüperparameetri sobivuse muutus võrreldes eelmise väärtusega. See tähendab, et kuigi ühe treeningu korral ei kasutatud kõiki treeningandmeid, kasutati kõiki treeningandmeid kolme treeningkorra jooksul. Kui iga kolme mudeli treeningu iteratiivsel hüperparameetrite otsingul olid leitud sobivad hüperparameetrid, arvutati leitud hüperparameetrite keskmised. Arvutatud hüperparameetrite abil loodi lõplik mudel. Lõpliku täpsuse saamiseks arvutati lõpliku mudeli ennustamise täpsus testandmete korral. (Seif, 2019)



Joonis 6. Andmestiku treening- ja testandmeteks jagamine, ristvalideerimine ja hüperparameetrite võrkotsing. Siniste värvidega on tähistatud mudeli ennustamise täpsuse leidmiseks kasutatavad andmed. Rohekama värviga on tähistatud valideerimise andmed.

Mudeli treenimise käigus leiti parima ennustustäpsusega matemaatiline funktsioon (nt lineaarne regressioon $y = w_i x_i$). Treenimisel leitavate mõõtmete osakaalud on mudeli parameetrid. Mitmetel masinõppemudelitel on ka hüperparameetrid, mis määratakse manuaalselt ja erinevad vastavalt algoritmile. Näiteks lineaarset regressiooni eristab kantregressioonist lisatud regularisatsiooniparameeter. (Zheng, 2015) Kuna igal algoritmil on erinevad hüperparameetrid ja neile määratavad väärtused, kasutati optimaalsete hüperparameetrite ja nende väärtuste leidmiseks võrkotsingut. Võrkotsingu abil optimaalsete hüperparameetrite leidmiseks kasutati sõnastikus määratud hüperparameetreid. Kõigi hüperparameetrite kombinatsioonide sobivus arvutati välja iga ristvalideerimise osas treenitud mudelite ennustamise täpsusi võrreldes. Näiteks LinearSVC algoritmi puhul leiti viie hüperparameetri optimaalseimad väärtused (lisa 8). Igale hüperparameetrile vastas mitu sobivat väärtust. Selle mudeli jaoks oli 48 erinevat hüperparameetrite kombinatsiooni. Iga kombinatsiooni treeniti kolme erineva valideerimise andmete hulga abil, st kokku treeniti LinearSVC lõpliku mudeli jaoks $48 * 3 = 144$ eraldi masinõppemudelit. Kuna kõigi kolme andmestiku korral toimiti samal viisil, treeniti kolme andmestikuga kokku 432 masinõppemudelit. Sobivate arvuliste väärtuste leidmine nõuab katsetamist ja eelteadmisi algoritmist. (Zheng, 2015) Uurimistöös vaadeldi parameetrite väärtusi väiksema vahemiku piires. Kui leiti, et optimaalne väärtus on vaadeldava vahemiku viimane väärtus, suurendati vaadeldava vahemiku väärtuseid. Sellist lähenemist nimetatakse manuaalseks võrkotsinguks. Manuaalne võrkotsing oli teiste võrkotsingutega võrreldes arvutuslikult kõige efektiivsem, kuigi eeldas hüperparameetrite mitmekordset muutmist ja igal korral mitmetunnise võrkotsingu käivitamist.

Masinõppemudelite loomisel hinnati masinõppemudeli keeleoskustaseme ennustamise täpsust testandmete korral skaalal 0–1, kus 1 tähistas veatut keeleoskustaseme ennustamise võimet (lisa 8). Täpsus leiti, kasutades valemit

$$t(\hat{y}_i, y_i) = \frac{1}{n} \sum_{i=0}^{n-1} \hat{y}_i = y_i,$$

kus n on tekstide arv, t leitud täpsus, $\hat{y}_i$ on i tekstile ennustatud keeleoskustase ja y_i on i on teksti tegelik keeleoskustase. (Scikit-learn'i kodulehekülg) Saadud tulemused koguti massiivi tulemuste hilisemaks visualiseerimiseks. Sisendandmete keeleoskustasemed olid kodeeritud diskreetseteks numbrilisteks väärtusteks. Kui mudel põhines regressioonil, ei olnud väljundandmed diskreetsed väärtused ja seetõttu ümardati saadud tulemused lähima diskreetse väärtuseni sarnaselt Vajjala ja Lõo regressioonil põhinevatele masinõppemudelitele (Vajjala, Lõo, 2014, lk 114).

4. TULEMUSTE ANALÜÜS

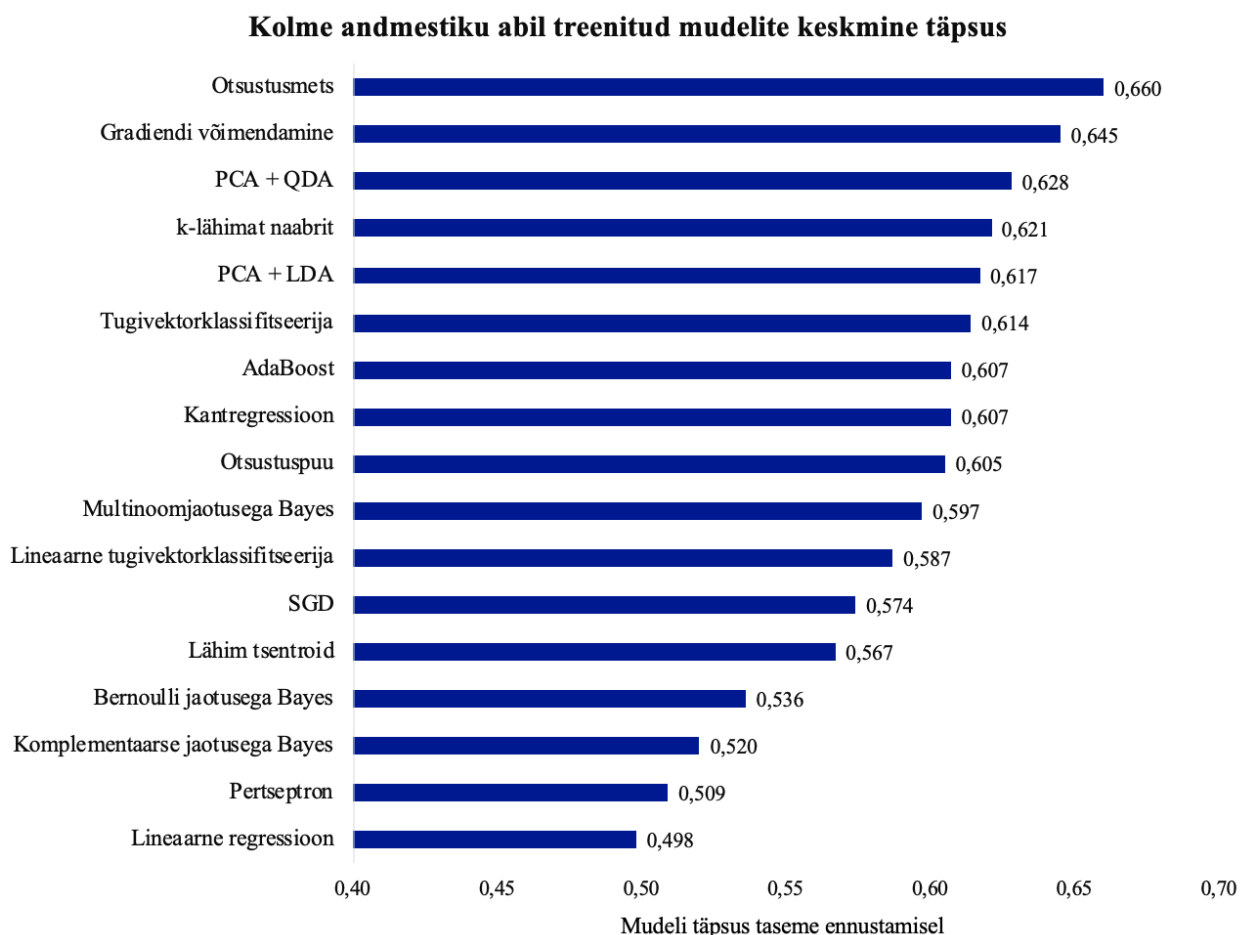
Uurimistöö empiirilises osas tehti 17 katset iga andmestikuga ning saadi 51 masinõppemudelit. Mudelite optimaalsed hüperparameetrid, täpsus testandmete klassifitseerimisel ja eksimismaatriksid on esitatud lisa (lisa 9). Eksimismaatriksid näitavad tekstide tegelikke keeleoskustasemeid ja ennustatud keeleoskustasemete jaotumist. Nende abil hinnati mudeli sobivust kvalitatiivselt. Tulemuste analüüsi esimeses alapeatükis kirjeldatakse kõige paremini keeleoskustaset ennustava algoritmi kindlaksmääramist kahel viisil. Esimesena võrreldakse kolme andmestiku abil treenitud mudelite keskmist keeleoskustaseme ennustamise täpsust. Teisena võrreldakse kõigi treenitud mudelite seast parimaid ennustamise täpsusi eraldi omavahel. Teises alapeatükis kirjeldatakse klassifitseerimiseks sobivaimaid andmestikke kahel viisil. Esimesena võrreldakse kolme andmestikuga treenitud mudelite keskmisi ennustamise täpsusi omavahel. Teisena leitakse, millise andmestikuga treenitakse 51 mudeli seast kõrgeima täpsusega masinõppemudelid.

4.1. Masinõppemudelite analüüs

Esimesena vaadeldakse iga mudeli keskmist ennustamise täpsust kolme erineva andmestikuga treenimisel. Keskmisi ennustamise täpsusi võrreldes leiti, millise mudeli tööpõhimõtte sobib sarnase andmestiku kasutamiseks. Parima keskmise ennustamise täpsusega mudel võib kasutamiseks sobida ka teiste mitmemõõtmeliste andmete korral, näiteks käändevormid või sõnaliikide ja sõnavormide järjendid. Parimad keskmise täpsusega algoritmid olid ansambelõppe algoritmid otsustusmets ja gradiendi võimendamine ning mõõtmete vähendamise algoritmid ruutfunktsiooni diskriminantanalüüs (joonisel PCA+QDA) ja lineaarne diskriminantanalüüs (joonisel PCA+LDA) (joonis 7). Parimate keskmise täpsusega algoritmide hulka kuulusid ka kaks andmepunktipõhiste algoritmide rühma algoritmi: k-lähima naabri algoritm ja tugivektorklassifitseerija. Kõige madalama keskmise ennustamise täpsusega oli regressiooni rühma lineaarse regressiooni mudelil. Madalad ennustamise täpsused olid samuti Bayesi algoritmide rühma Bernoulli jaotusega Bayesi algoritmil ja komplementaarse jaotusega Bayesi algoritmil ning neurovõrkude rühma algoritmil pertseptron.

Ansambelõppe algoritmide hulka kuulusid otsustusmets, gradiendi võimendamine ja Adaboost. Nende seast andsid täpseima tulemuse otsustusmetsa ja gradiendi võimendamise mudelid. Adaboosti kasutamisel on probleemiks ülesobituvus. Kuna otsustusmets kasutab

100–150 eraldi otsustuspuud, on otsustusmets võrdlemisi resistentne ülesobituvuseks. Tulemustest võib järeldada, et ennustamise täpsust suurendab ülesobituvuse võimaluse vähendamine ning et andmestikus esineb suures hulgas võõrväärtusi ja segavaid tunnuseid, mille arvesse võtmine vähendab mudeli täpsust. Mõõtmete vähendamise rühma kuuluvad ruutfunktsiooni kasutav diskriminantanalüüs ja lineaarne diskriminantanalüüs. Nende seast saavutas kõrgema ennustamise täpsuse ruutfunktsiooni kasutav diskriminantanalüüs. Mõlema algoritmi kasutamisel vähendati suurel määral väheinformatiivsete mõõtmete mõju, kuna juhitud masinõppele lisaks kasutati ka juhendamata masinõppe mõõtmete vähendamist PCA algoritmi abil. Mõõtmete vähendamise algoritmide kõrge täpsus ennustamisel kinnitab, et andmestikku iseloomustab suur tõenäosus ülesobituvuseks. Eemaldades väheinformatiivsed tunnused, saavutatakse täpsem tulemus, kuna treenitud mudel ei arvesta suurema tõenäosusega juhuslikult muutuvaid tunnuseid.



Joonis 7. Masinõppemudelite keskmine ennustamise täpsus kolme andmestiku korral. Mudeli täpsust hinnati skaalal 0–1, kus 1 tähendab täiuslikku keeleoskustaseme ennustamise täpsust.

Andmepunktipõhiste algoritmide rühmast uuriti nelja algoritmi: lineaarne tugivektorklassifitseerija, mittelineaarne tugivektorklassifitseerija, k-lähima naabri algoritm ja lähima tsentroidi algoritm. Andmepunktipõhiste algoritmide seast saavutasid kõrgemad ennustamise täpsused k-lähima naabri algoritm ja mittelineaarne tugivektorklassifitseerija. Radiaalbaasfunktsiooni kasutatav mittelineaarne tugivektorklassifitseerija algoritm (lisa 9) saavutas tunduvalt parema ennustamise täpsuse kui lineaarne tugivektorklassifitseerija. Radiaalbaasfunktsioon loob andmepunktide eristamiseks mittelineaarselt uue tasandi ja selle kasutamisel saavutatud kõrge ennustamise täpsus näitab, et keeleoskustasemeid eristavad otsustuspiirid on keeruliselt määratavad. K-lähima naabri algoritm saavutas kõrgema tulemuse kui lähima tsentroidi algoritm. Viie kuni kaheksa lähima andmepunkti klassi põhjal määrava k-lähima naabri algoritmi tugevus on keeruliste otsustuspiiride arvestamine ning selle algoritmi kõrgem ennustamise täpsus kinnitab, et otsustuspiirid on keeruliselt määratavad. Andmepunktipõhiste algoritmide nõrkadeks külgedeks on tundlikkus vöörväärtuste osas ja madal ennustustäpsus kõrgema mõõtmete arvu korral. Töös kasutatavad 17 mõõdet teevad andmepunktipõhiste algoritmide korral kauguse leidmise punktide vahel keeruliseks ja ebatäpseks, mistõttu saavutavad ansambelõppe ja mõõtmete vähendamise algoritmid kõrgema täpsuse.

Otsustuspuu algoritmi kasutamisel saavutati teiste masinõppemudelitega võrreldes keskmine ennustamise täpsus. Otsustuspuu algoritmi optimaalne sügavus oli 4–8 taset, mis tähendab, et 17 tunnuse seast ennustati keeleoskustase kasutades vaid 4–8 kõige informatiivsemat tunnust. Otsustuspuu algoritm sobib hästi uuritava andmestiku alusel keeleoskustaseme ennustamiseks, kuid selle kasutamisel saadakse madal ennustamise täpsus, kuna ülesobituvusest hoidumiseks ei kasutata kõiki tunnuseid (tabel 4). Otsustuspuud rakendava, kuid kõiki tunnuseid vaatleva otsustusmetsa algoritmi kasutamisel saavutati tunduvalt kõrgem ennustamise täpsus. Naiivsete Bayesi algoritmide kasutamisel saavutati madalad ennustamise täpsused. Naiivsete Bayesi algoritmide rakendamisel on vajalik eeldus, et andmestikus esinevate tunnuste vahel ei esine korrelatsiooni. Selle kasutamise eelised on vähene tundlikkus vöörväärtuste suhtes. Keeruliselt määratavad otsustuspiirid ja sõnaliikide korreleeruvad tunnused (lisa 3) tingisid algoritmide madalad keskmised ennustamise täpsused. Naiivsete Bayesi algoritmide seast saavutati kõrgemad ennustamise tulemused kasutades multinoomjaotust ja Bernoulli jaotust.

Tabel 4. Masinõppemudelite peamised puudujäägid

Mudelite rühm	Puudused treenimisel
Regressioon	Seoste ülelihtsustamine, korreleeruvad tunnused
Andmepunktipõhised	Võõrväärtuste esinemine, eriti B1 ja B2 keeleoskustasemete vahelisel otsustuspiiril, ebatäpne tulemus suurema mõõtmete arvu korral
Otsustuspuud	Korreleeruvad tunnused, madala sügavusega klassifitseerija, kõigi tunnuste mitte kasutamine
Bayesi algoritmid	Ebasobilik eeldus, korreleeruvad tunnused
Neurovõrgud	Ebapiisav kogus treeningandmeid, ülesobituvus, kalduvus ennustada vaid ühte keeleoskustaset
Mõõtmete vähendamine	Keerulisemate hüpertasandi funktsioonide mitte arvesse võtmine
Ansambelõpe	Ülesobituvus Adaboosti algoritmi korral

Neurovõrkude algoritmide pertseptron ja stohhastiline gradientlaskumine (joonisel SGD) kasutamisel saavutati madalad ennustamise täpsused. Neurovõrkude treenimiseks on vajalik suur treeningandmete kogus ja need on tundlikud võõrväärtuste esinemise suhtes. Pertseptroni ja stohhastilise gradientlaskumise eksimismaatriksites (lisa 9) oli näha, et mudelid ennustasid suure tõenäosusega B1 taset või C1 taset. Pertseptroni eksimismaatrikseid teiste masinõppemudelite eksimismaatriksitega võrreldes oli märgata, et neurovõrkude rakendamisel keeleoskustaseme ennustamiseks oli vajalik suurem kogus treeningandmeid. Stohhastilise gradientlaskumise kasutamisel saavutati parem keeleoskustaseme ennustamise täpsus, sest otsustuspiiri leidmisel ei piirdutud lineaarse funktsiooniga. Lineaarne regressioon ennustas keeleoskustasemeid kõige ebatäpsemalt, sest on kõige algelisem ennustusalgoritm. Ridge'i kulufunktsiooni kasutatav kantregressioon andis tunduvalt kõrgema ennustamise täpsuse, kuna vähendas ülesobituvust ja tundlikkust võõrväärtuste suhtes.

Mudelite keskmisi keeleoskustasemete ennustamise täpsusi võrreldi, et leida, millised masinõppemudelid võivad sobida sarnastele tekstide iseloomustavatele andmestikele. Tekstiliste tunnuste väärtused varieerusid keeleoskustasemete lõikes suurel määral, sest keeleõppijate taust ja sellest tulenev keelekasutus on väga erinev. Seetõttu tuleb

automaathindamise masinõppemudeli loomisel arvestada, et treenitav masinõppemudel ei oleks tundlik võõrväärtuste suhtes. Samuti tuleb arvestada mittelineaarsete otsustuspiiride ja mitmemõõtmelise andmestikuga. Kolmest põhjusest tulenevalt võivad sarnaste andmestike treenimisel kõrge ennustamise täpsuse anda ansambelõppe ja mõõtmete vähendamise algoritmid. Keeleoskustaset võivad potentsiaalselt määrata ka neurovõrkude masinõppemudelid, kui on võimalik kasutada suuremas koguses treeningandmeid.

Kõigi treenitud mudelite parimaid ennustamise täpsusi võrreldi, et leida milline masinõppemudel on kõige sobivam sõnaliike kirjeldava andmestiku abil keeleoskustaseme ennustamiseks. Kõrgeima ennustamise täpsusega masinõppemudelit on võimalik rakendada automaathindamise rakendusse. Uurimistöös treeniti kõrgeima ennustamise täpsusega mudel kasutades gradiendi võimendamise algoritmi ja andmestikuks sõnaliigiti sõnatüvede osakaalu tekstisõnadest. Gradiendi võimendamise mudeli eksimismatriksil on näha kalduvus ennustada tekstile keeleoskustasemeks B1 taset (joonis 8). B2 tasemega tekstide keeleoskustasemeks ennustati võrdlemisi kõrge (38%) tõenäosusega B1 keeleoskustase. Masinõppemudel ennustas kõige täpsemini A2 ja C1 keeleoskustasemeid.

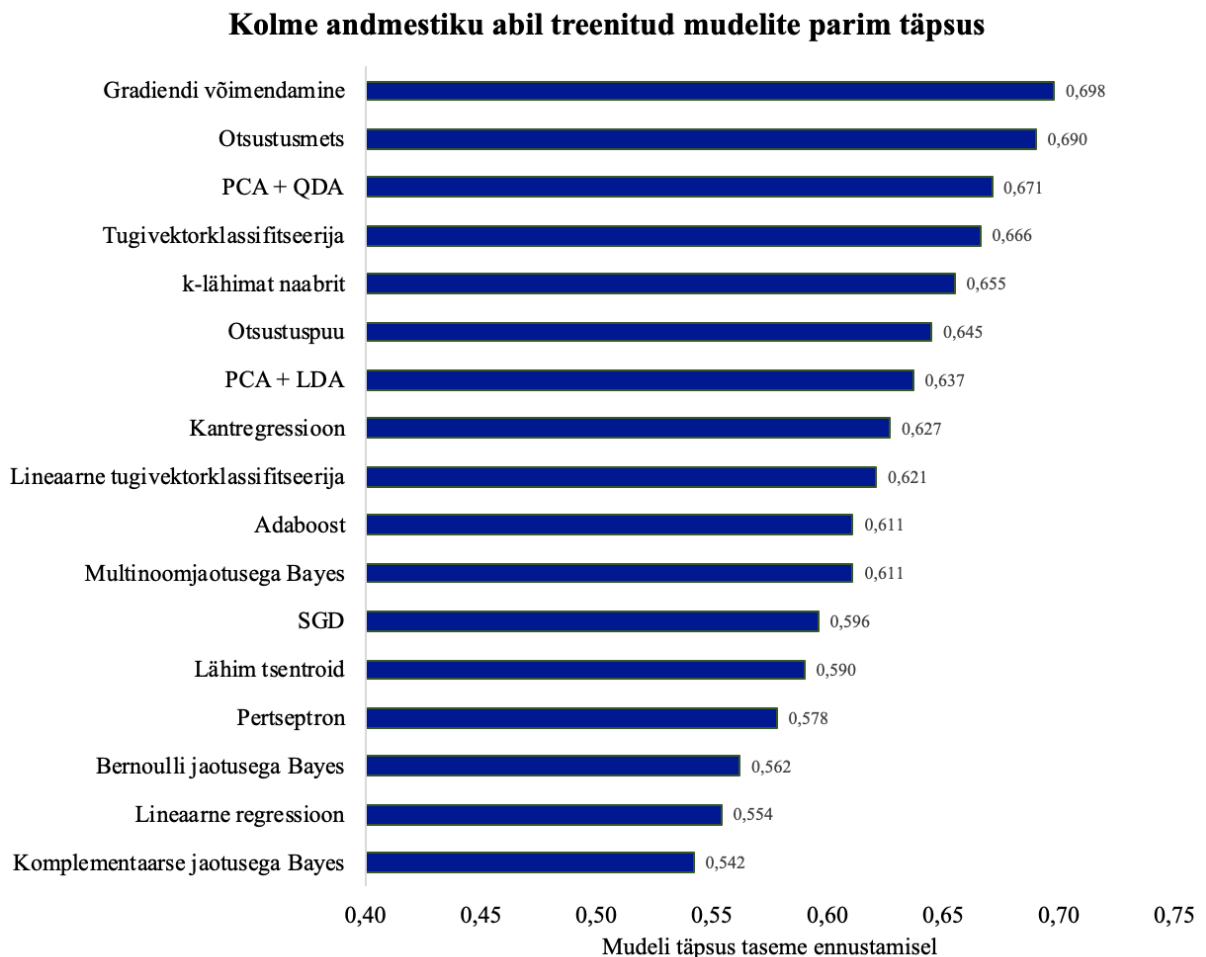
Gradiendi võimendamise mudeli eksimismatriks

Teksti keeletase	A2	0.68	0.27	0.06	0.0
	B1	0.05	0.85	0.09	0.01
	B2	0.03	0.39	0.52	0.05
	C1	0.0	0.2	0.22	0.58
		A2	B1	B2	C1
		Ennustatud keeletase			

Joonis 8. Gradiendi võimendamise masinõppemudeli eksimismatriks, sõnatüvede sõnaliikide osakaal tekstist andmestik.

Kõrgeim saavutatud täpsus oli 0,698 ehk 69,8% (joonis 9). Parima ennustamise täpsusega masinõppemudelid olid sarnaselt keskmise ennustamise täpsuse mudelite tulemustega ansambelõppe ja mõõtmete vähendamise algoritmide mudelid. Erinevalt keskmise ennustamise täpsusega mudelite tulemustest saavutati parim ennustamise täpsus

tugivektorklassifitseerijat kasutades. Parimaid ja keskmisi ennustamise täpsusi võrreldes on märgatav, et parima ennustamise täpsusega masinõppemudelid leiavad mittelineaarse otsustuspiiri või kasutavad ansambelõpet. Mudeli jaoks kõige informatiivsema andmestiku treeningandmete kasutamisel oli väikseim mõju lineaarsel diskriminandi analüüsil (2%) ja ruutfunktsiooni diskriminantanalüüsil (4,3%).

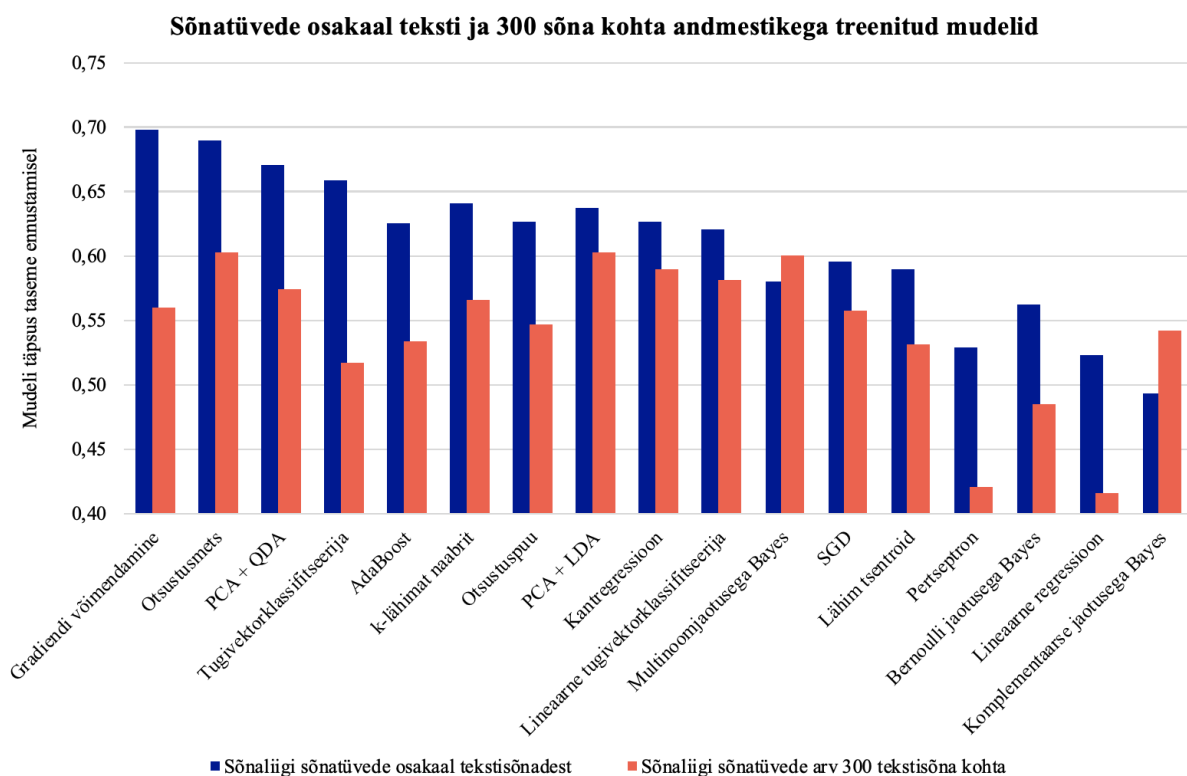


Joonis 9. Masinõppemudelite parim ennustamise täpsus kolme andmestiku korral. Mudeli täpsust hinnati skaalal 0–1, kus 1 tähistab täiuslikku keeleoskustaseme ennustamise täpsust.

4.2. Andmestike analüüs

Üks uurimistöö hüpoteese oli, et kõige täpsem mudel saadakse treenides andmestikuga, kus on leitud iga sõnaliigi erinevate sõnatüvede osakaal, võttes igas tekstis arvesse 300 tekstisõna. Selleks, et võrrelda kuidas mõjutab tulemust see, kui sõnatüvede ja tekstisõnade suhtarvused arvutatakse terve teksti asemel 300 tekstisõna kohta, võrreldi kahe andmestiku abil treenitud mudelite täpsusi keeleoskustaseme ennustamisel (joonis 10). Peale kahe masinõppemudeli saavutati igal mudelil kõrgem täpsus, kui mudel oli treenitud andmestikuga,

kus oli leitud iga sõnaliigi sõnatüvede osakaal terve teksti tekstisõnade kohta. Võrreldes treenitud mudelite keskmise ennustamise täpsusi saab järeldada (tabel 5), et püstitatud hüpotees ei leidnud kinnitust. Põhjused, miks andmestik, milles võeti arvesse iga sõnaliigi eripäraste sõnatüvede osakaalu terve teksti kohta, andis täpsemaid ennustamise tulemusi, võisid olla järgnevad: tekstide kasutamine kirjutatud ja mitte kokku liidetud kujul ning tekstipikkuse mõju andmestiku kõrgemale informatiivsusele.

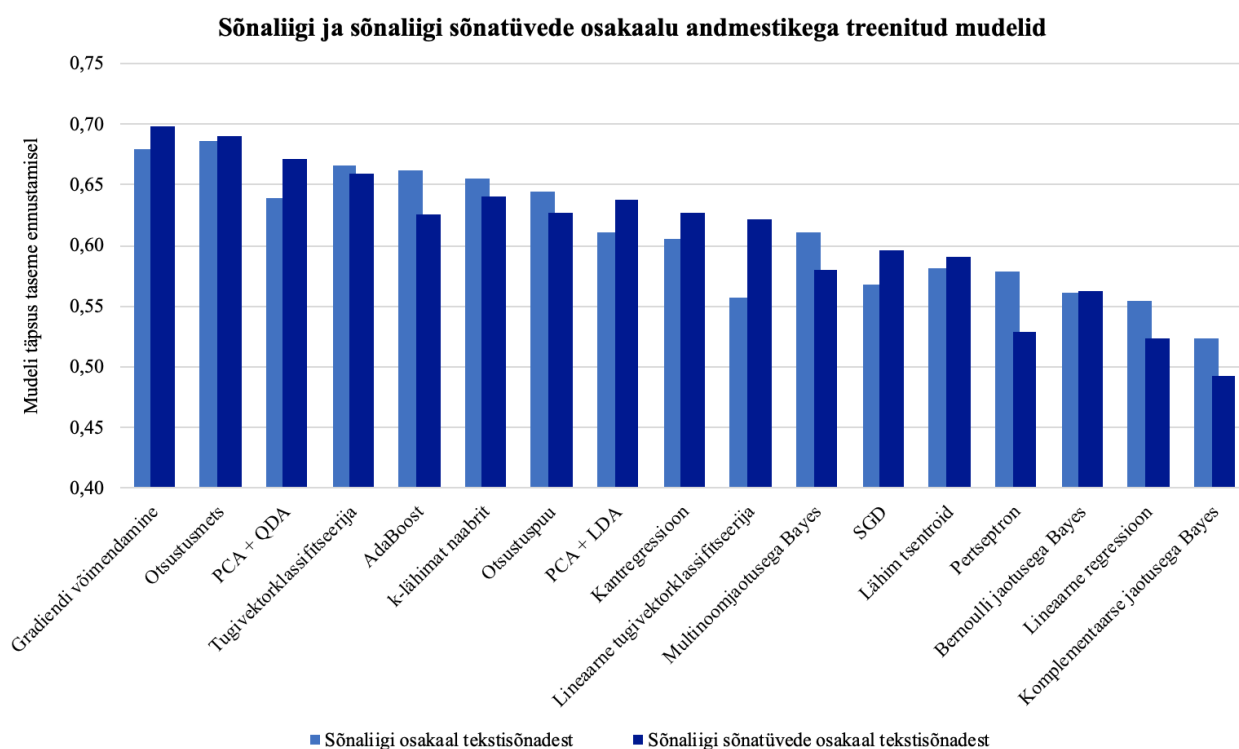


Joonis 10. Masinõppemudelite ennustamise täpsused, treenitud andmetega terve teksti kohta ja 300 sõna kohta. Mudeli täpsust hinnati skaalal 0–1, kus 1 tähistab täielikku keeleoskustaseme ennustamise täpsust.

Tabel 5. Eri andmestike abil treenitud masinõppemudelite keskmine keeleoskustaseme ennustamise täpsus

Andmestik	Andmestikuga treenitud mudelite keskmine täpsus
Sõnaliigi osakaal sõnade ehk tekstisõnade hulgas	0,613
Erineva tüvega sõnade ja kõigi tekstisõnade suhtarv sõnaliigiti	0,614
Sõnaliigi sõnatüvede osakaal 300 tekstisõna kohta	0,546

Lisaks teksti pikkusele võrreldi töös sõnatüvede arvestamise mõju treenitud masinõppemudeli ennustamise täpsusele. Selleks võrreldi kahe andmestikuga treenitud mudelite ennustamise täpsust. Ühes andmestikus leiti sõnaliigi osakaal sõnade ehk tekstisõnade hulgas, teises andmestikus leiti erineva tüvega sõnade ja kõigi tekstisõnade suhtarv sõnaliigiti. Mõlemas andmestikus leiti osakaal teksti tekstisõnade kohta. Kahe andmestikuga treenitud mudelite keskmised erinesid üksteisest vähe (tabel 5), kuid erinevus ilmnis kõige tõhusamate masinõppemudelite korral (joonis 11). Kolm kõige tõhusamat masinõppemudelit olid kõrgema ennustamise täpsusega, kui need treeniti andmestikuga, kus leiti erineva tüvega sõnade ja kõigi tekstisõnade suhtarv sõnaliigiti. Andmestik, kus võetakse arvesse sõnaliikide sõnatüvesid on olulisema praktilise väärtusega ning on seetõttu eelistatum Eesti vahekeele korpuse automaatse keeleoskustaseme hindamise rakenduses kasutamiseks.



Joonis 11. Masinõppemudelite ennustamise täpsused, treenitud andmetega sõnaliigi osakaalu ja sõnaliigi sõnatüvede osakaalu kohta. Mudeli täpsust hinnati skaalal 0–1, kus 1 tähistab täielikku keeleoskustaseme ennustamise täpsust. Masinõppemudelid on järjestatud parima täpsusega masinõppemudelite järjestuses.

KOKKUVÕTE

Uurimistöös uuriti, kuidas masinõppe abil sõnaliikide kasutust analüüsid on võimalik kõige tõhusamalt automaatselt keeleoskustaset määrata. Sõnaliikide kasutuse uurimiseks kasutati allikatena 985 Eesti vahekeele korpuse teksti. Kõige tõhusama masinõppemudeli leidmiseks võrreldi 17 algoritmi. Töö üks eesmärk oli leida parim sõnaliikide andmete esitamise viis, et saavutada parim tulemus tekstide taseme hindamisel. Selleks võrreldi kolmel viisil esitatud andmeid: sõnaliigi osakaal tekstisõnades, sõnatüvede osakaal tekstisõnades sõnaliigiti ning sõnaliigi sõnatüvede osakaal 300 tekstisõnas. Püstitati hüpotees, et kõrgeim ennustamise täpsus saavutatakse, treenides masinõppemudeleid andmestikuga, kus arvestatakse sõnaliigi sõnatüvede osakaalu 300 tekstisõna kohta. Töös selgus, et kõrgeima ennustamise täpsusega mudelid saadi, kasutades kaht andmestikku: sõnaliigi osakaal tekstisõnades (61,3%) ja sõnatüvede osakaal tekstisõnades sõnaliigiti (61,4%). Andmestike võrdluses ei mõjutanud sõnatüvede arvesse võtmine oluliselt keskmist keeleoskustaseme ennustamise täpsust. Kõige madalama keskmise ennustamise täpsusega (54,6%) mudelid saadi, kui kasutati andmestikku, kus arvestati sõnaliigi sõnatüvede osakaalu 300 tekstisõna kohta. Madalast ennustamise täpsusest saadi kinnitust, et püstitatud hüpotees ei osutunud tõeseks. Masinõppemudelite ennustamise täpsusi eraldi analüüsid selgus, et kõige tõhusamate masinõppemudelite treenimisel saavutatakse kõrgem täpsus andmestikuga, kus leiti sõnatüvede osakaal tekstisõnades sõnaliigiti. Andmestik, kus võeti arvesse sõnaliikide sõnatüvesid, omab suurimat praktilist väärtust automaatse keeleoskustaseme määramise rakenduse loomiseks.

Töö teine eesmärk oli leida kõige tõhusam mudel tekstide keeleoskustaseme määramiseks. Püstitati hüpotees, et kolme andmestiku lõikes kõige kõrgema keskmise ennustamise täpsusega on otsustusmetsa algoritm. Töös leiti, et kolm kõige kõrgema keskmise ennustamise täpsusega masinõppemudelit on järjestatult otsustusmets, gradiendi võimendamine ja peakomponentide analüüs kombineeritult ruutfunktsiooni diskriminandi analüüsiga. Hüpoteesile, et kolme andmestiku lõikes kõige kõrgema keskmise ennustamise täpsusega on otsustusmetsa masinõppemudel, saadi kinnitust. Madalad ennustamise täpsused saadi neurovõrkude kasutamisel, sest vähem kui tuhat teksti on ebapiisav juhendamata masinõppe mudeli treenimiseks. Parima eraldi arvestatava ennustamise täpsusega masinõppemudelit olid järjestatult gradiendi võimendamine, otsustusmets ja peakomponentide analüüs kombineeritult ruutfunktsiooni diskriminandi analüüsiga. Töös jõuti järeldusele, et suurim praktiline väärtus oli gradiendi võimendamisel. Uuritud 17 tunnust

kasutades saavutati skaalal 0–1 parimaks ennustamise täpsuseks 0,698 ehk 69,8%. Saavutatud täpsus on tunduvalt väiksem Vajjala ja Lõo treenitud mudelist, mille täpsus oli 79% ja mis oli treenitud 78 tunnusel. Siiski tõstab sõnaliikide abil keeleoskustaseme ennustamise tõhusamaks muutmine potentsiaalselt ka tulevikus loodavate suurema tunnuste arvuga masinõppemudelite täpsust.

Uurimistöö eesmärk saavutati, sest töös leiti parim esitusviis sõnaliikide kasutust kirjeldavate andmete jaoks ja kõige sobivam masinõppemudel. Masinõppealased ettepanekud edasiseks uurimiseks on suurema andmestikuga juhendamata masinõppe rakendamine ja täiendavate ansambelõppe mudelite katsetamine. Keelealased ettepanekud edasiseks uurimiseks on tekstide täpsem sõnestamine ja suurema arvu tunnuste kasutamine. Eesti keele automaatse keeleoskustaseme rakenduse täpsuse suurendamine eeldab täiendavat uurimist, kuidas kasutada lisaks grammatilistele tunnustele ka sõnavara ja sidususe tunnuseid, tekstitasandi ja sõnatasandi keelevigu. Kõrge informatiivsusega tunnusteks võivad osutuda sõnatasandi keelevead sõnaliigiti, mida pole veel kasutatud masinõppemudeli treenimiseks. Eesti keele erinevus indoeuroopa keeltest loob võimaluse uurida teiste keelte keeleoskustaseme eristamisel vähetähtsate tunnuste rakendamist ja loodetavasti jätkatakse erinevate tunnuste uurimist eesti keele jaoks kasutatava keeletehnoloogia arendamiseks.

RESÜMEE

Kirjaliku keeleoskuse taseme määramine sõnaliikide põhjal

Automaatse keeleoskustaseme hindamine on oluline eksamite hindajatele ja keeleõppijatele tagasiside saamiseks ning samuti eesti keele jaoks kasutatava keeletehnoloogia arendamiseks. Kuna eesti keel erineb suurema kõnelejaskonnaga indoeuroopa keeltest, ei ole loodud keeletehnoloogia kasutamiseks rakendatav ja samaväärselt arendatud. Uurimistööd tehes sooviti leida kõige tõhusamad masinõppemudelid tekstide tasemete hindamiseks sõnaliikide ja sõnaliigiti sõnatüvede põhjal.

Sõnaliikide ja sõnaliigiti sõnatüvede erinevaid esitamise viise võrreldi, kasutades kolme andmestikku: sõnaliigi osakaal tekstisõnades, sõnatüvede osakaal tekstisõnades sõnaliigiti ning sõnatüvede arv sõnaliigiti 300 tekstisõna kohta. Iga andmestiku abil treeniti 17 masinõppemudelit ning saadud 51 masinõppemudeli täpsust hinnati skaalal 0-1, kus 0 vastas alati valesti ja 1 veatule teksti keeleoskustaseme ennustamisele. Töös selgus, et kõrgeim keskmine täpsus saavutati otsustusmetsa algoritmi abil (*decision tree*). Parima ennustamise täpsusega masinõppemudel treeniti gradiendi võimendamise (*gradient boosting*) algoritmi abil, kasutades treeningandmeteks sõnatüvede osakaalu tekstisõnadest sõnaliigiti. Kõrgeim saavutatud ennustamise täpsus oli 69,8%. Parimate tulemustega masinõppemudelid kuulusid ansambelõppe ja mõõtmete vähendamise algoritmide rühmadesse. Töös selgus, et masinõppemudeli loomisel on oluline vähendada ülesobituvust. Vastavalt treeningandmete kasutamisele saadi erinevad tulemused ennustamise täpsuste keskmistel: iga sõnaliigi osakaalu tekstisõnades andmetega 61,3%, sõnatüvede osakaalu tekstisõnades sõnaliigiti andmetega 61,4% ning sõnatüvede arvu sõnaliigiti 300 tekstisõna kohta andmetega 54,6%. Kolm parimat masinõppemudelit saavutasid kõrgeima täpsuse kasutades treeningandmeteks sõnatüvede osakaalu tekstisõnadest sõnaliigiti. Uurimistöös selgus, et andmete osakaalu võtmine 300 sõne kohta oli vähem informatiivne kui osakaalu võtmine kogu teksti sõnede kohta. Tulemustest järeldati, et sõnaliikide andmeid kasutades saab tõhusalt teksti keeleoskustaset määrata, kuid mitmete väheuuritud tunnuste analüüsimine ja suurema treeningandmete kogusega on võimalik saavutada täpsem ennustusmudel. Järeldus, et teistes keeltes väheinformatiivsed sõnaliikide tunnused on edukalt rakendatavad eesti keelele suunatud masinõppes, on oluline eesti keele jaoks kasutatava keeletehnoloogia arendamiseks.

ABSTRACT

Using grammatical tagging for automated language proficiency level prediction

Developing automated language level prediction tools is essential for giving feedback to Estonian language learners, to ensure accurate assessment at language exams and for developing language technology for the Estonian language. Automated language level prediction of the Indo-European languages has been widely researched, but as Estonian differs significantly from those languages, different features such as position of speech (POS) tags might prove to be very informative. For analysing the importance of POS tags in automated language level prediction on CEFR scale, three types of data were used to train Machine Learning models. The data about the ratio of POS tags to the word tokens, the ratio of distinct lemmas per POS tags to the word tokens and the ratio of distinct lemmas per POS tags to 300 words were used. 958 texts that were used were collected by Tallinn University Estonian Learner Language Corpus (EVKK) from learners of Estonian as a second language. For finding the best Machine Learning model, 17 models were trained for each type of data. The best performing models were trained with data about distinct lemmas per POS tags to the word tokens. The highest prediction accuracy of 69,8% was achieved with data about the ratio of distinct lemmas per POS tags to the word tokens and using Gradient Boosting algorithm. The larger implication of the research was the contribution to the development of the Estonian Learner Language Corpus (EVKK) automated language level prediction tool.

KASUTATUD MATERJALID

Allkivi, K. 2016. C1-tasemega eesti keele õppijate ja emakeelekõnelejate kirjaliku keelekasutuse võrdlus verbialguliste tetragrammide näitel. Kättesaadav: <http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/lahivordlusi/article/view/LV26.02/399>. (24.02.2020).

Allkivi-Metsoja, K. 2019. Keeleoskustasemete statistiline mudeldamine eesti keele kui teise keele kirjalike tekstide automaatseks hindamiseks [Doktoritöö metodoloogiline plaan]. Tallinn: Tallinna Ülikool.

Bacallado, S. 2012. Classification, LDA. Kättesaadav: <https://web.stanford.edu/class/stats202/content/lec9.pdf>. (12.12.2020).

Behzadi, F. 2015. Natural Language Processing and Machine Learning: A Review. Kättesaadav: https://www.academia.edu/16382956/Natural_Language_Processing_and_Machine_Learning_A_Review. (21.02.2020).

Brainlee, J. 2019 A Tour of Machine Learning Algorithms. Kättesaadav: <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>. (12.01.2020).

Breiman, L. 1998a. Arcing classifiers (with discussion). *The Annals of Statistics*, 26, lk 801–849.

Breiman, L. 1998b. Combining predictors [teadustöö aruanne]. Berkeley: California ülikool.

Durgesh, K., Bhambhu, L. 2009. Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, 1, lk 1–6.

Eesti vahekeele korpus. Mis on EVKK. Kättesaadav: http://evkk.tlu.ee/wwwdata/what_is_evk. (28.01.2020).

Eslon, P. 2007. Õppijakeele korpused ja keeleõpe. Kättesaadav: http://evkk.tlu.ee/wwwdata/kogumik2007/korpusekogumik_pille.pdf. (03.01.2020).

Eslon, P. 2014 Eesti vahekeele korpus. *Keel ja Kirjandus*, 6, lk 436–451.

Euroopa Nõukogu 2001. Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Freund, Y., Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1, lk 148–156.

- Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. Kättesaadav: <https://statweb.stanford.edu/~jhf/ftp/trebst.pdf> (13.01.2020).
- Friedman, J., Hastie, T., Tibshirani, R. 2000. Additive Logistic Regression: a statistical view of boosting. *The Annals of Statistics*. nr 2, lk 337–407.
- Friedman, J., Hastie, T., Tibshirani, R. 2009 The elements of statistical learning. Kättesaadav: www.web.stanford.edu/~hastie/ElemStatLearn/. (12.01.2020).
- Goyal, P., Pandey, S., Jain, K. 2018. Deep Learning for Natural Language Processing. New York: Apress.
- Haridus- ja Teadusministeerium. 2014. Eesti elukestva õppe strateegia 2020. Kättesaadav: <https://www.hm.ee/sites/default/files/strateegia2020.pdf> (24.02.2020).
- Haridus- ja Teadusministeerium. 2018a. „Eesti keele arengukava 2011–2017“ elluviimise. Kättesaadav: https://books.google.ee/books?id=7z_KAgAAQBAJ&pg=PT54&lpg=PT54&dq=type+token+ratio+100+word+chunks&source=bl&ots=XjK48yofvH&sig=ACfU3U0OH03cw2DbXGeZND19SZHolhqiMQ&hl=et&sa=X&ved=2ahUKEwiBpL2x-fbmAhU7wMQBHf2OAEwCnoECBUQAQ#v=onepage&q=type%20token%20ratio%20100%20word%20chunks&f=false. (24.02.2020).
- Hyafil, L., Rivest, R. L. 1976. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5, lk 15–17.
- Katz, A., Tao, X., Maltby, H. Multivariate Normal Distribution. Kättesaadav: <https://brilliant.org/wiki/multivariate-normal-distribution/>. (11.12.2020).
- Kennedy, G. 1998. An introduction to corpus linguistics. New York: Longman.
- Kerge, K. 2012. Kas õpikuteksti mõõdetavad tunnused on kõnekad? Kättesaadav: http://www.emakeeleselts.ee/omakeel/2012_2/OK_2012-2_03.pdf. (24.02.2020).
- Kerge, K., Pajupuu, H., Altrov, R. 2007. Tekst, kontekstuaalsus ja kultuur. Kättesaadav: https://www.researchgate.net/publication/261724512_Tekst_kontekstuaalsus_ja_kultuur_Text_contextuality_and_culture. (24.02.2020).
- Korjus, K., Hebart, M. N., Vicente, R. 2016. An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. Kättesaadav: <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0161788&type=printable> (21.02.2020).

- Kossinski, J. 2018. Masinõppel rajaneva tarkvararakenduse loomine keeleoskustaseme ennustamiseks [bakalauresetöö]. Tallinn: Tallinna Ülikool.
- Kotsiantis, S. B. 2007. Supervised Machine Learning: A Review of Classification Techniques. Kättesaadav: [https://datajobs.com/data-science-repo/Supervised-Learning-\[SB-Kotsiantis\].pdf](https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf). (21.02.2020).
- Lõo, K. 2012. The Relationship between Lexical Richness and Estonian Learners Language Proficiency Levels [seminaritöö käsikiri]. Tübingen: Eberhard Karls Universität Tübingen.
- Lõo, K., Vajjala, S. 2014- Automatic CEFR Level Prediction for Estonian Learner Text. Proceedings of the third workshop on NLP for computer-assisted language learning, lk 113–127.
- Martinez, A. M., Kak, A. C. 2001 PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23, lk 228–23.
- Meunier, F. 2013. Computer tools for the analysis of learner corpora. Learner English on Computer, lk 19–37. London: Longman.
- Molnar, C. 2019. Interpretable Machine Learning. Kättesaadav: <https://christophm.github.io/interpretable-ml-book/>. (02.01.2020).
- Prabhakaran, S. How Naive Bayes Algorithm Works? (with example and full code). Kättesaadav: <https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/>. (02.01.2020).
- Raschka, S. 2014. Linear Discriminant Analysis. Kättesaadav: http://sebastianraschka.com/Articles/2014_python_lda.html. (11.12.2019).
- Raschka, S. 2015- Single-Layer Neural Networks and Gradient Descent. Kättesaadav: https://sebastianraschka.com/Articles/2015_singlelayer_neurons.html. (25.02.2020).
- Schapire, R. E., Freund, Y. 2012. Boosting: Foundations and Algorithms. Kättesaadav: https://doc.lagout.org/science/0_Computer%20Science/2_Algorithms/Boosting_%20Foundations%20and%20Algorithms%20%5BSchapire%20%26%20Freund%202012-05-18%5D.pdf. (21.02.2020).
- Schapire, R. E., Singer, Y. 1998. Improved boosting algorithms using confidence-rated predictions. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1, lk 80–91.

- Scikit-Learn. 1. Supervised learning. Kättesaadav: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. (02.01.2020).
- Seeger, M. 2012. Pattern Classification and Machine Learning. Kättesaadav: https://www.cc.gatech.edu/grads/j/jwang774/pcml_notes.pdf. (24.02.2020).
- Seif, G. 2019. Why and How to do Cross Validation for Machine Learning. Kättesaadav: <https://towardsdatascience.com/why-and-how-to-do-cross-validation-for-machine-learning-d5bd7e60c189>. (18.01.2020).
- Selinker, L. 1992. Interlanguage. Error Analysis. Perspectives on Second Language Acquisition. London: Longman, lk 31–54.
- Shmeleva, A. 2010. Vene õppijakeele korpuse koostamise põhimõtted. Kättesaadav: http://evkk.tlu.ee/2010/2010_slaidid. (10.12.2020).
- Smola, A., Vishwanathan, S. V. N. 2008. Introduction to Machine Learning. Ühendkuningriigid: Cambridge University Press.
- Zheng, A. 2015. Evaluating Machine Learning Models. Kättesaadav: <https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/ch04.html>, (11.12.2020).
- Taiwo, O. A. 2010. Types of Machine Learning Algorithms, New Advances in Machine Learning. Kättesaadav: <http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms>. (11.12.2020).
- Talib, R., Hanif, M. K., Ayesha, S., Fatima, F. 2016. Text Mining: Techniques, Applications and Issues. (IJACSA) *International Journal of Advanced Computer Science and Applications*, 11, lk 414–418.
- Tono, Y. 2000. A corpus-based analysis of interlanguage development: analysing pos tag sequences of EFL learner corpora. In PALC'99: Practical Applications in Language Corpora, pages 323–340.
- Verlinde, P., Maître, G., Mayoraz, E. 1998. Decision fusion using a multi-linear classifier. Kättesaadav: https://www.researchgate.net/publication/37432695_Decision_Fusion_Using_a_Multi-Linear_Classifier. (03.02.2020).
- Vyatkina, N. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, 96, lk 576–598.

Taiwo, O. A. 2010. Vasakul tugivektorklassifitseerija mudel, kus klasse eraldav otsustuspiir on väikese marginaaliga. Paremal kõrgema ennustamise täpsusega mudel, kus otsustuspiir on võimalikult suure marginaaliga [Joonis 1]. Kättesaadav: <http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms> (11.12.2020).

Schapire, R. E., Freund, Y. 2012. Vasakul diagrammil eraldi otsustuspuu klassifitseerijate treenimine. Treenimist alustatakse vasakpoolsest otsustuspuust. Paremal joonisel on näidatud eraldi treenitud klassifitseerijate lisamine lõplikusse mudelisse [Joonis 3]. Kättesaadav: https://doc.lagout.org/science/0_Computer%20Science/2_Algorithms/Boosting_%20Foundations%20and%20Algorithms%20%5BSchapire%20%26%20Freund%202012-05-18%5D.pdf. (21.02.2020).

Kotsiantis, S. B. 2007. Juhendatud masinõppemudeli loomise töökäik [Joonis 4]. Kättesaadav: [https://datajobs.com/data-science-repo/Supervised-Learning-\[SB-Kotsiantis\].pdf](https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf). (21.02.2020).

Scikit-learn. Andmestiku treening- ja testandmeteks jagamine, ristvalideerimine ja hüperparameetrite võrkotsing. Siniste värvidega on tähistatud mudeli ennustamise täpsuse leidmiseks kasutatavad andmed. Rohekama värviga on tähistatud valideerimise andmed. [Joonis 6]. Kättesaadav: https://scikit-learn.org/stable/_images/grid_search_cross_validation.png. (21.02.2020).

Lõo, K., Vajjala, S. 2014. Eestikeelsete tekstide 10 kõige informatiivsemat tunnust keeleoskustasemete eristamisel. Kaldkriipsuga on tunnuste nimetustes tähistatud jagamistehet [Tabel 1]. Automatic CEFR Level Prediction for Estonian Learner Text. Proceedings of the third workshop on NLP for computer-assisted language learning.

Brainlee, J. (2019) Empiirilises osas kasutatavad algoritmid jagatuna sarnasuse alusel rühmadesse [Tabel 2]. Kättesaadav: <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>. (12.01.2020).

EVKK Kuueastmelises süsteemis (A1, A2, B1, B2, C1, C2) on ekspertide hinnang antud kokku 987 tekstile [Tabel 3]. Kättesaadav: <http://evkk.tlu.ee>. (28.01.2020).

Lisa 1. Lingvistiliseks analüüsiks kasutatavad tunnused

Tunnuste rühm	Tunnuste alamrühm	Tunnuste näited	Kasutatavad rakendused
Grammatika	Morfoloogilised tunnused	Nt sõnaliik, käändevorm, kõneviis, ajavorm, arv, tegumood; nimi-, tegu- ja asesõna tüüp; sõnaliikide ja -vormide järjendid	Rakendused: StanfordNLP või EstNLTK morfoloogia- ja süntaksianalüsaator, Klastreleidja
	Süntaktilised tunnused	Nt lauseliige, kaassõna ja sidendi tüüp, lauseliikmete järjendid	
	Morfosüntaktilised tunnused	Nt sõnaliikide/sõnavormide ja lauseliikmete kombinatsioonid, nende järjendid	
Sõnavara ja sidusus	Leksikaalsed tunnused	Nt liitsõnad, sõnavara mitmekesisus (erinevate sõnade ja kõigi tekstisõnade suhtarv, ka sõnaliigiti), keerukus (põhisõnavara jm sõnade suhtarv), tihedus (sisu- ja funktsioonisõnade suhtarv)	Rakendused: StanfordNLP või EstNLTK lemmatiseerija ja morfoloogiaanalüsaator
	Diskursiivsed tunnused	Nt sidusvahendid (sidesõnad, teatud ase- ja mäarsõnad)	
	Semantilised tunnused	Nt sõna semantiline klass ja semantiline roll lauses, pärisnimed ja nende liigid	Rakendused: EstNLTK Wordneti liidestus, nimega üksuste tuvastaja
		Nt tähevead, algustähe vead, kokku-lahkukirjutamine	

Täpsus	Sõnatasandi veatunnused		Rakendus: EstNLTK õigekirjakorrektor
	Tekstitasandi veatunnused	Nt sõnajärg, puuduolevad ja üleliigsed komad, grammatilised seosed - ühildumine ja sõltumine	Rakendused: eesti keele grammatikakorrektor, sõnajärje vealeidja

(Allkivi-Metsoja, 2019)

Lisa 2. Algoritmide klassifitseerimine sarnasuse alusel

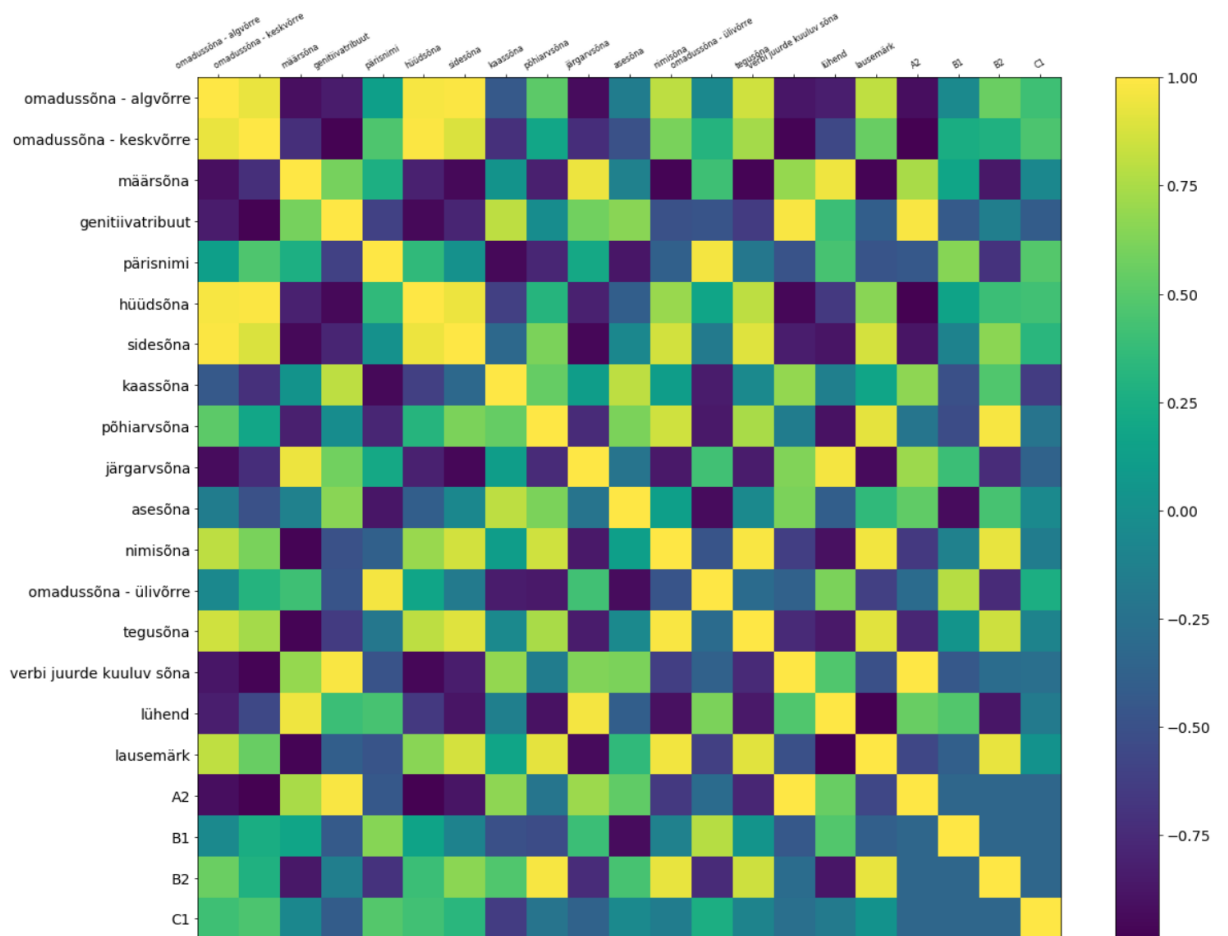
Algoritmide rühm	Algoritmid
Regression	Hariliku vähemruutude meetodi regressioon Lineaarne regressioon Logistiline regressioon Sammregressioon Mitmemõõtmelise adaptiivse regressiooni splainid (MARS) Lokaalselt määratud korrelatsioonivälja hajutamine (LOESS)
Andmepunktipõhised	k-lähimat naabrit (kNN) Õppiv vektori kvantimine (LVQ) Ise-organiseeriv kaart (SOM) Lokaalselt kaalutud õpe (LWL) Tugivektorklassifitseerija (SVM)
Regularisatsioon	Kantregressioon Least Absolute Shrinkage and Selection Operator (LASSO) <i>Elastic Net</i> Väiksema nurga regressioon (LARS)
Otsustuspuud	Klassifitseerimise ja regressiooni otsustuspuu (CART) <i>Iterative Dichotomiser 3</i> (ID3) C4.5 ja C5.0 <i>Chi-squared Automatic Interaction Detection</i> (CHAID) <i>Decision Stump</i> M5 Tinglikud otsustuspuud
Bayesi algoritmid	Naiivne Bayes Gaussi jaotusega naiivne Bayes

	Multinoomjaotusega naiivne Bayes Averaged One-Dependence Estimators (AODE) <i>Bayesian Belief Network</i> (BBN) Bayesi võrgustik (BN)
Klasterdusalgoritmid	k-keskmised k-mediaanid Ootuste maksimeerimine (EM) Hierarhiline klasterdamine
Seostepõhised algoritmid	Apriori algoritm Eclat' algorithm
Neurovõrgud	Pertseptron Mitmekihilised pertseptronid (MLP) Tagasilevi Stohhastiline gradientlaskumine Hopfieldi võrgustik Radiaalbaasfunktsiooni võrgustik (RBFN)
Süvaõpe	Konvolutsiooniline närvivõrk (CNN) Rekurrentne närvivõrk (RNNs) <i>Long Short-Term Memory Networks</i> (LSTMs) Kihilised autokooderid Sügav Boltzmanni masin (DBM) <i>Deep Belief Networks</i> (DBN)
Mõõtmete vähendamise algoritmid	Peakomponentide analüüs (PCA) Peakomponendi regressioon (PCR) Osaline vähimruutude regressioon (PLSR) Sammoni kaardistamine

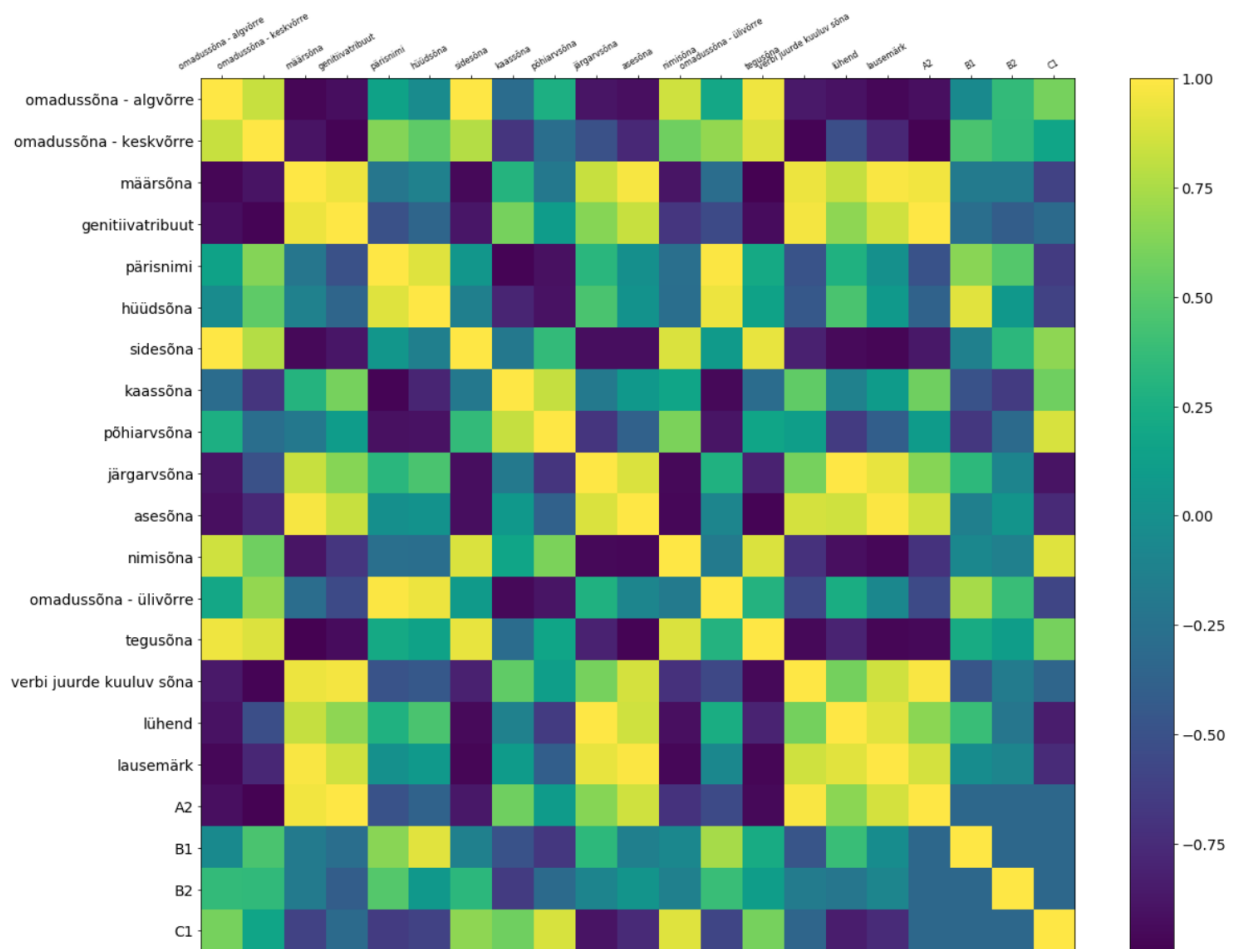
	Mitmemõõtmeline skaleerimine (MDS) <i>Projection Pursuit</i> Lineaarne diskriminandi analüüs (LDA) Mitmefunktsiooniline diskriminandi analüüs (MDA) Ruutfunktsiooni diskriminandi analüüs (QDA) Paindlik diskriminandi analüüs (FDA)
Ansambelõpe	Buustimine <i>Bootstrapped Aggregation (Bagging)</i> AdaBoost Kaalutud keskmine (<i>Blending</i>) Kihiline üldistamine (<i>Stacking</i>) Gradiendi buustimine (GBM) Gradiendi buustimisega regressiooni otsustuspuud (GBRT) Otsustusmets

(Brainlee, 2019)

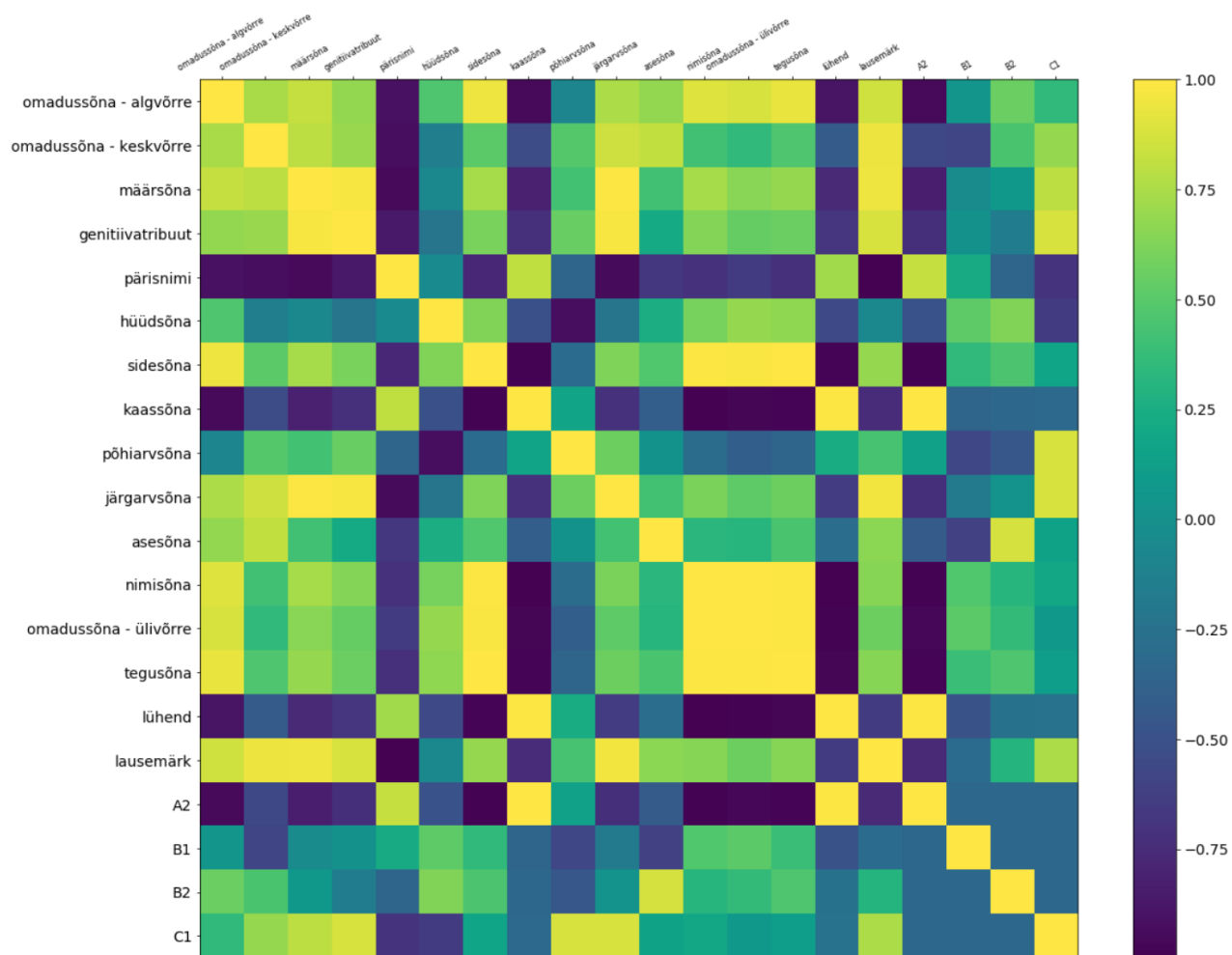
Lisa 3. Andmete ja keeleoskustasemete korrelatsioonimaatriksid



Sõnaliikide osakaalu tekstisõnadest ja keeleoskustasemete korrelatsioonimaatriks

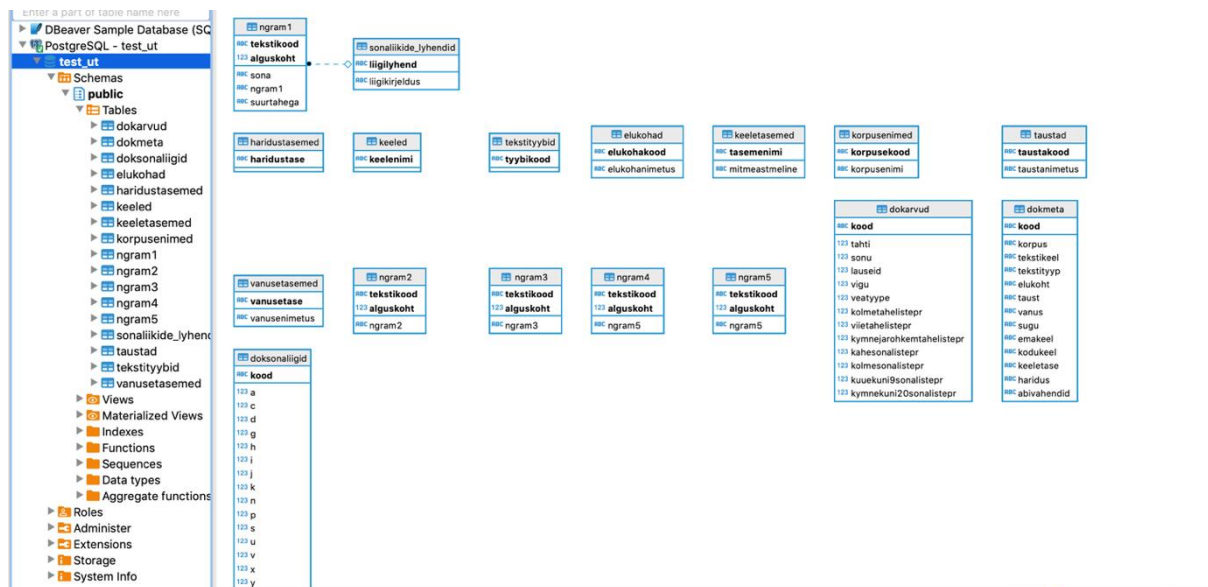


Sõnaliikide sõnatüvede osakaalu tekstisõnadest ja keeleoskustasemete korrelatsioonide maatriks



Sõnaliikide eri sõnatüvede osakaal 300 tekstisõna kohta ja keeleoskustasemete korrelatsioonide maatriks

Lisa 4. Andmebaasi skeemi taastamine



Korpuse andmete kopeerimise jaoks relatsioonilise andmebaasi skeemi taastamine

Lisa 5. Tekstide andmed

„dokmeta“ andmetabeli esimesed sissekanded, kuhu on sisestatud kõigi tekstide tekstikoodid ja võõrvõtmed, mis refereerivad teistele andmetabelitele.

textcode	corpuscode	textlanguage	texttype	residence	socback
doc_100636852915_item	cFOoRQekA	eesti	essee	idaviru	op
doc_100636852916_item	cFOoRQekA	eesti	muu	idaviru	op
doc_100636852917_item	cFOoRQekA	eesti	essee	idaviru	op

age	gender	motherlanguage	homelanguage	proficiencylevel	education	useddictionary
kuni18	naine	vene	vene	B	pohi	ei
kuni18	naine	vene	vene	B	pohi	ei
kuni18	naine	vene	vene	B	pohi	ei

„ngram1“ andmetabeli andmed tekstikoodiga doc_100636852915_item viie esimese sõna kohta

tekstikood	sona	sonaliik	sonanrtekstis	suurtahega
doc_100636852915_item	suvel	S	1	Suvel
doc_100636852915_item	ma	P	2	ma
doc_100636852915_item	lugesin	V	3	lugesin
doc_100636852915_item	läbi	D	4	läbi
doc_100636852915_item	ajalehe	S	5	ajalehe

```

index,tekstikood,keeletase,A,C,D,G,H,I,J,K,N,O,P,S,U,V,X,Y,Z
0,doc_104580264036_item,C1,22.0,1.0,30.0,6.0,3.0,0.0,38.0,5.0,3.0,2.0,60.0,69.0,0.0,92.0,0.0,0.0,55.0
1,doc_104580264038_item,C1,27.0,2.0,51.0,7.0,0.0,0.0,31.0,5.0,2.0,3.0,45.0,81.0,2.0,79.0,0.0,0.0,66.0
2,doc_104580264040_item,C1,30.0,3.0,54.0,0.0,0.0,0.0,38.0,2.0,2.0,1.0,61.0,63.0,0.0,86.0,0.0,0.0,56.0
3,doc_104580264043_item,B2,32.0,2.0,43.0,5.0,8.0,0.0,48.0,1.0,1.0,0.0,72.0,56.0,0.0,102.0,0.0,0.0,61.0
4,doc_104580264044_item,C1,29.0,0.0,59.0,8.0,1.0,0.0,23.0,5.0,1.0,3.0,54.0,74.0,3.0,79.0,0.0,3.0,61.0
5,doc_104580264045_item,C1,34.0,1.0,48.0,12.0,6.0,0.0,32.0,10.0,4.0,0.0,65.0,78.0,0.0,101.0,1.0,0.0,78.0
6,doc_104580264046_item,B2,31.0,1.0,48.0,0.0,3.0,0.0,35.0,4.0,1.0,6.0,84.0,52.0,2.0,91.0,0.0,0.0,53.0
7,doc_104580264047_item,C1,18.0,4.0,61.0,10.0,8.0,0.0,26.0,6.0,3.0,9.0,36.0,96.0,2.0,66.0,0.0,0.0,54.0

```

csv-vormingus salvestatud tekstide andmed

Pandas teigi DataFrame objekti salvestatud andmed, kus veerud on tähistatud sõnaliikide lühenditega

	keeletase	A	C	D	G	H	I	J	K	N	O	P	S	U	V	X	Y	Z
0	C1	22.0	1.0	30.0	6.0	3.0	0.0	38.0	5.0	3.0	2.0	60.0	69.0	0.0	92.0	0.0	0.0	55.0
1	C1	27.0	2.0	51.0	7.0	0.0	0.0	31.0	5.0	2.0	3.0	45.0	81.0	2.0	79.0	0.0	0.0	66.0
2	C1	30.0	3.0	54.0	0.0	0.0	0.0	38.0	2.0	2.0	1.0	61.0	63.0	0.0	86.0	0.0	0.0	56.0
3	B2	32.0	2.0	43.0	5.0	8.0	0.0	48.0	1.0	1.0	0.0	72.0	56.0	0.0	102.0	0.0	0.0	61.0
4	C1	29.0	0.0	59.0	8.0	1.0	0.0	23.0	5.0	1.0	3.0	54.0	74.0	3.0	79.0	0.0	3.0	61.0

Allikas: EVKK andmete alusel joonis pandas andmetabelist

Suvel ma lugesin läbi ajalehe "....", kus oli väga huvitav artikkel ühest koolist, väikesest alevikust Võru juures - Sõmerpalust.

Selles koolis on väga sõbralik ja tähelepanelik kollektiiv.

Viis aastat tagasi kooliõpilased ja õpetajad koos tegid väikeses toad koolikeskuses talveaeda.

See oli raske töö, sest nad ise pidid tegema suuri toru kuuma veega, et ruumis oleks soe, ja hiiglasuuri lampe, et olik valgus.

Ja nüüd see on väga ilus koht.

Toa keskel kasvavad erinevad troopilised taimid, õitsevad ilusad lilled, aga piki seima seisavad suured puurid, kus elavad papagoid.

Nurgas seisab veel üks suur puur - seal elab madu - suur tiigri püüton.

Üks kord talvel, kui Sõmerpalus ei oli elektrit ja sooja vett 5 - 6 päevat, õpetajad ja õpilased pidid viima kooli kodust pudeleid sooja veega ja sojendama nendega madu, et ta ei sureks.

See näitab, kui vastuvad õpilased koolis on, ja kooli kollektiiv on nagu suur pere.

Mulle meeldib see atikkel, sest ma sain teada, et väikeses linnas, kus, kui ma tean, on väga raske elu, inimesed katsuvad luua midagi meeldiv ja hea nende jaoks, et oleks natukene rõmsaks.

Unicode-vormingus txt-failina salvestatud tekst.

Allikas: EVKK

Teksti kaheksa esimese sõna sõnatüved ja nende vastavad sõnaliigid

Teksti sõna	Sõnatüvi	Sõnaliigi tähistus	Sõnaliik
KONTROLLTÖÖ	kontrolltöö	S	nimisõna
EESTI	Eesti	H	pärisnimi
KULTUURILOOST	kultuurilugu	S	nimisõna
Eestlastele	eestlane	S	nimisõna
muinasusundi	muinasusund	S	nimisõna
üldiseloostusMis	üldiseloostusmis	P	pärisnimi
on	olema	V	tegusõna
alepõllundus	alepõllundus	S	nimisõna

Allikas: EVKK andmetel koostatud tabel

```
,filename,A,C,D,G,H,I,J,K,N,O,P,S,U,V,X,Y,Z,tekstikood,keeletase
15,doc_767882159235_item.txt,4.966887417218543,0.3311258278145696,5.298013245033113,0.0,2.3178807947019866,0.0,1.3245033112582782
,0.6622516556291391,1.3245033112582782,1.6556291390728477,3.3112582781456954,19.86754966887417,0.3311258278145696,6.6225165562913
92,0.0,0.0,2.6490066225165565,doc_767882159235_item.txt,B1
47,doc_278690626920_item.txt,4.5662100456621,0.0,3.6529680365296797,0.0,2.28310502283105,0.0,1.8264840182648399,0.456621004566209
96,1.8264840182648399,0.9132420091324199,3.1963470319634704,19.17808219178082,0.0,8.67579908675799,0.0,0.0,2.7397260273972597,doc
278690626920_item.txt,B1
```

Saadud andmed salvestatuna csv-vormingus.

	A	C	D	G	H	I \
keeletase						
A2	3.500865	0.102078	6.400843	0.523020	6.097827	0.048292
B1	5.606527	0.363128	9.631555	0.496810	2.613037	0.100993
B2	6.649234	0.505739	10.160858	0.485066	2.662833	0.093441
C1	6.532387	0.551386	9.702766	0.448107	3.328006	0.035850

	J	K	N	O	P	S \
keeletase						
A2	3.858675	0.829104	3.639875	0.809192	12.839032	24.756519
B1	6.069354	1.308849	1.211323	0.667682	12.252064	22.799308
B2	6.554804	1.582113	0.932648	0.803802	10.787643	23.969068
C1	6.513691	1.789502	3.237074	1.346954	10.085556	24.638637

	U	V	X	Y	Z	total
keeletase						
A2	0.025128	17.410729	0.015753	1.456404	17.686664	183.788018
B1	0.041237	19.021345	0.022304	0.263180	17.531305	269.851582
B2	0.040079	18.237302	0.021182	0.409482	16.104706	466.271930
C1	0.072181	16.742760	0.028794	0.386586	14.559763	896.705426

Sõnaliikide osakaal keeleoskustasemete järgi.

	A	C	D	G	H	I \
keeletase						
A2	2.767266	0.089335	4.870864	0.423468	4.157090	0.046188
B1	4.237699	0.317218	6.481319	0.216174	1.874457	0.082755
B2	5.237684	0.425073	6.291121	0.219821	1.716539	0.077515
C1	5.010585	0.425776	5.382342	0.154394	1.686623	0.030913

	J	K	N	O	P	S \
keeletase						
A2	1.400650	0.675151	3.280021	0.760857	3.743430	18.267962
B1	1.736418	0.994079	1.051498	0.513736	3.082385	15.411165
B2	1.404015	1.225253	0.819600	0.579928	2.748861	15.486369
C1	1.074276	1.273182	2.576702	0.738358	2.033320	13.373839

	U	V	X	Y	Z
keeletase					
A2	0.025128	8.741599	0.012596	1.124713	2.610716
B1	0.037524	8.638989	0.021348	0.169470	2.097817
B2	0.034098	8.221670	0.020688	0.325409	1.694632
C1	0.054852	6.831264	0.021385	0.213244	1.145244

Sõnaliikide sõnatüvede osakaal tekstisõnadest keeleoskustasemete kohta.

	A	C	D	G	H	I \
keeletase						
A2	0.790452	0.875168	0.760972	0.809658	0.681733	0.956427
B1	0.755851	0.873570	0.672926	0.435125	0.717348	0.819416
B2	0.787712	0.840498	0.619152	0.453178	0.644629	0.829565
C1	0.767037	0.772192	0.554722	0.344548	0.506797	0.862273

	J	K	N	O	P	S \
keeletase						
A2	0.362987	0.814314	0.901136	0.940268	0.291566	0.737905
B1	0.286096	0.759506	0.868058	0.769431	0.251581	0.675949
B2	0.214196	0.774441	0.878788	0.721482	0.254816	0.646098
C1	0.164926	0.711473	0.795997	0.548169	0.201607	0.542799

	U	V	X	Y	Z
keeletase					
A2	1.000000	0.502081	0.799630	0.772254	0.147609
B1	0.909957	0.454173	0.957136	0.643933	0.119661
B2	0.850760	0.450816	0.976669	0.794686	0.105226
C1	0.759931	0.408013	0.742672	0.551609	0.078658

Unikaalsete sõnatüvede osakaal erinevate sõnaliikide ja keeleoskustasemete korral.

	A	C	D	G	H	I \
keeletase						
A2	0.677196	0.108534	0.259751	0.006426	23.435205	0.056157
B1	1.633174	0.106209	0.456968	0.022330	19.269981	0.129692
B2	2.123981	0.264663	0.482029	0.018584	15.176865	0.137605
C1	1.922395	0.301290	0.664034	0.040800	12.789439	0.044640

	J	K	N	O	P	S \
keeletase						
A2	0.050454	0.307023	0.379258	0.123192	0.098243	10.391187
B1	0.097374	0.005350	0.054705	0.189438	0.068551	17.077204
B2	0.101380	0.011163	0.109308	0.213731	0.304513	16.367474
C1	0.091563	0.014641	0.707796	0.316627	0.190434	15.816784

	U	V	Y	Z
keeletase				
A2	0.023662	0.511922	3.188777	12.295490
B1	0.078600	1.109225	0.723339	15.035117
B2	0.073127	1.143594	1.111701	18.543494
C1	0.062627	0.991929	1.121785	21.222581

Unikaalsete sõnatüvede osakaal 300 sõna kohta keeleoskustasemete korral.

Lisa 6- Andmete töötlemise kood

```
: conn_string = "host="+ PGHOST +" port="+ "5432" +" dbname="+ PGDATABASE +" user=" + PGUSER \
+" password="+ PGPASSWORD
conn=psycopg2.connect(conn_string)
cursor = conn.cursor()

# Defineeri andmete laadimise funktsioon
def load_data(schema, table):
    sql_command = "SELECT * FROM {}.{};".format(str(schema), str(table))

    # Lae SQL käsu abil vajalik tabel
    data = pd.read_sql(sql_command, conn)
    return (data)

# Lae alla ngram1 andmetabel
ngram1 = load_data('public', 'ngram1')

# Grupeeri andmed tekstide põhjal. Grupeeri uuesti iga teksti sõ
grouped_by_doc = ngram1.groupby(['tekstikood', 'ngram1'], as_index=False).sona.count()

# Lae alla dokmeta andmetabel
dokmeta = load_data('public', 'dokmeta')

# Liida dokmeta ja ngram1 andmetabelid tekstikoodi põhjal kokku.
df_outer = pd.merge(grouped_by_doc, dokmeta, left_on='tekstikood', right_on='kood', how='outer')

# Vali välja ainult huvipakkuvad veerud
ngramcount_and_level = df_outer[['tekstikood', 'ngram1', 'sona', 'keeletase']]
```

```
# Filtreeri välja kõigist tekstidest 958 teksti nelja uuritava keeletasemega
df = ngramcount_and_level[ngramcount_and_level['keeletase'].isin(['A2', 'B1', 'B2', 'C1'])]

table = pd.pivot_table(df, values='sona', index=['tekstikood', 'keeletase'],
                        columns='ngram1')

# Täida puuduvad väljad 0-ga
new_table = table.fillna(0)

# Salvesta andmed CSV failivormis
new_table.to_csv('puhtad_1_andmed.csv')
```

```
: import os
from estnltk import Text
import pandas as pd

# Valmista ette tühi andmetabel
dataframe = pd.DataFrame()

# Otsi tekste kaustast dokumendid
directory = 'dokumendid'

# Vaata läbi iga tekst
for filename in os.listdir(directory):
    failisisu = open(filename).read()

    # Muuda tekstisisu estnltk Text objektiks
    text = Text(failisisu)

    # Leia iga sõnatüve sõnavorm
    text_data = text.get.word_texts.lemmas.postags.postag_descriptions.as_dataframe

    # Vali välja esimene estnltk poolt ennustatav sõnavorm
    text_data['postags'] = text_data['postags'].apply(lambda x: x[0])

    # Eemalda andmetabelist korduvalt esinevad sõnatüved
    grouped_data = text_data.groupby(by='postags', as_index=False).agg({'lemmas': pd.Series.nunique})

    # Arvuta sõnatüvede osakaal teksti sõnadest
    words = len(text.words)
    grouped_data['lemmas'] = grouped_data.lemmas/words*100
    grouped_data = grouped_data.rename(columns={'lemmas': 'lemmas_percentage'})

    # Lisa andmetabelisse rida, kus veerg filename tähistab teksti
    grouped_data['filename'] = filename
    grouped_data = grouped_data.pivot(index='filename', columns='postags', values='lemmas_percentage')
    dataframe = dataframe.append(grouped_data)

# Täida puuduvad väljad 0-ga ja salvesta CSV formaadis
dataframe = dataframe.fillna(0)
dataframe.to_csv('postags_lemmas_data.csv')
```

```

# Lae alla mõlemad andmetabelid
lemmas_data = pd.read_csv('postags_lemmas_data.csv')
level_data = pd.read_csv('puhtad_voll_andmed.csv')

# Jäta alles ainult tekstikoodi ja keeletaseme veerud
level_data = level_data.drop(['A', 'C', 'D', 'G', 'H', 'I', 'J', 'K', 'N', \
                              'O', 'P', 'S', 'U', 'V', 'X', 'Y', 'Z'], 1)

# Lisa keeletaseme tabeli tekstikoodi veerule failiformaadi tähistus
level_data.tekstikood = level_data['tekstikood'] + ".txt"
level_data = level_data.drop('index', 1)

# Liida andmetabelid failinimede põhjal
df_outer = pd.merge(lemmas_data, level_data, left_on='filename', right_on='tekstikood', how='outer')

# Filtreeri välja neli keeletaset ja salvesta CSV formaadis
data = df_outer[df_outer['keeletase'].isin(['A2', 'B1', 'B2', 'C1'])]
data.to_csv('postags_lemmas_levels_data.csv')

```

```

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

# Keeletasemete kodeermine numbrilisteks väärtusteks
labelencoder = LabelEncoder()
table.iloc[:,0] = labelencoder.fit_transform(new_table.iloc[:,0])

# Keeletasemete ja sõnavormide üksteisest lahutamine
features = table.loc[:, "A":"Z"]
target_var = table.loc[:, "keeletase"]

# Sõnavormide osakaalu arvutamine kogu teksti sõnadest
features['total'] = features.sum(axis=1)
features = features.apply(lambda x: (x/features['total'])*100)

# Andmete jagamine treening- ja testandmeteks
X_train, X_test, y_train, y_test = \
    train_test_split(features.loc[:, "A":"Z"], target_var, train_size = 0.7, random_state=1111)

```

Allikas: Scikit-learni juhiste alusel loodud programm

Lisa 7. Sõnaliikide tähistused

Tähistus	Sõnaliik	Näide
A	omadussõna - algvõrre (adjektiiv - positiiv), nii käänduvad kui käändumatud	kallis või eht
C	omadussõna - keskvoorre (adjektiiv - komparatiiv)	laiem
D	määrsõna (adverb)	kõrvuti
G	genitiivatribuut (käändumatu omadussõna)	balti
H	pärisnimi	Edgar
I	hüüdsõna (interjektsioon)	tere
J	sidesõna (konjunktsioon)	ja
K	kaassõna (pre/postpositsioon)	kaudu
N	põhiarvsõna (kardinaalnumeraal)	kaks
O	järgarvsõna (ordinaalnumeraal)	teine
P	asesõna (pronoomen)	see
S	nimisõna (substantiiv)	pikim
U	omadussõna - ülivõrre (adjektiiv - superlatiiv)	pikim
V	tegusõna (verb)	lugema
X	verbi juurde kuuluv sõna, millel eraldi sõnaliigi tähistus puudub	plehkt
Y	lühend	USA

z	ausemärke	-, /, ...
---	-----------	-----------

Lisa 8. Masinõppemudelite treenimise kood

```
def benchmark(clf):
    print('=' * 80)
    print("Mudeli treenimine: ")
    print(clf)
    t0 = time()
    clf.fit(X_train, y_train)
    train_time = time() - t0
    print("Treeningu aeg: %0.3fs" % train_time)

    t0 = time()
    # Ennusta keeletase ja ümarda regressiooni mudeli ennustused
    pred = clf.predict(X_test)
    pred = [abs(round(x)) for x in pred]
    test_time = time() - t0
    print("Testi aeg: %0.3fs" % test_time)

    score = metrics.accuracy_score(y_test, pred)
    print("Täpsus: %0.3f" % score)

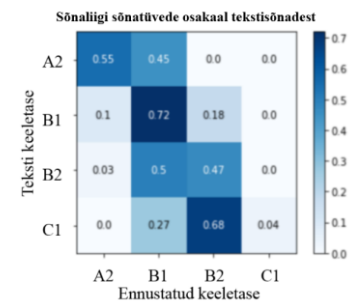
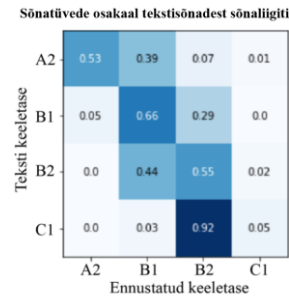
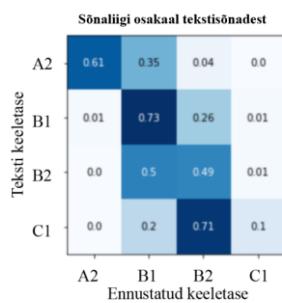
    print()
    clf_descr = str(clf).split('(')[0]
    return clf_descr, score, train_time, test_time

results = []
```

```
print('=' * 80)
print("LinearSVC")
lsvc_param_grid = {
    'penalty': ['l1', 'l2'],
    'loss': ['hinge', 'squared_hinge'],
    'dual': [False, True],
    'multi_class': ['crammer_singer', 'ovr'],
    'C': [2, 10, 20, 32, 150, 500, 2000],
    'max_iter': [2000, 2500, 5000]
}
lsvc = LinearSVC()
lsvc_search = GridSearchCV(lsvc, lsvc_param_grid, cv=3, return_train_score=True)
lsvc_search.fit(X_train, y_train)
results.append(benchmark(lsvc_search.best_estimator_))
```

Allikas: Scikit-learn'i juhiste alusel loodud programm

Lisa 9. Katsete eksimismatriksid



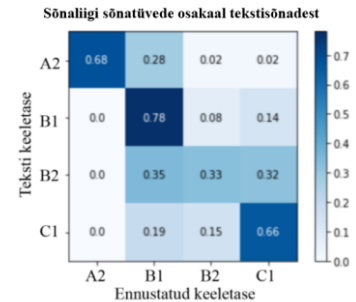
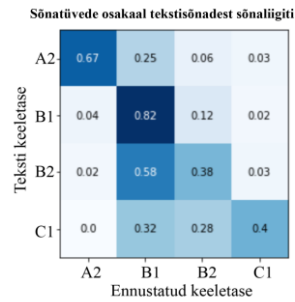
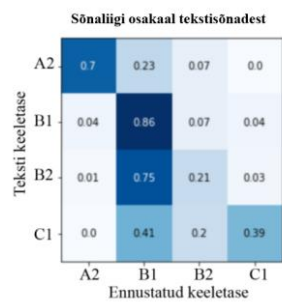
Täpsus

0,554

0,523

0,416

Lineaarse regressiooni eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.

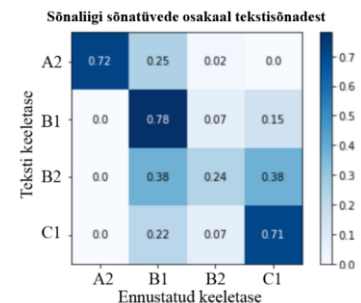
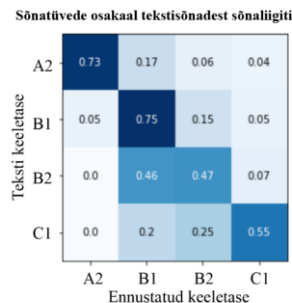
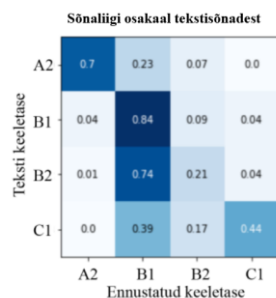


Täpsus 0,605

0,627

0,590

Kantregressiooni eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus

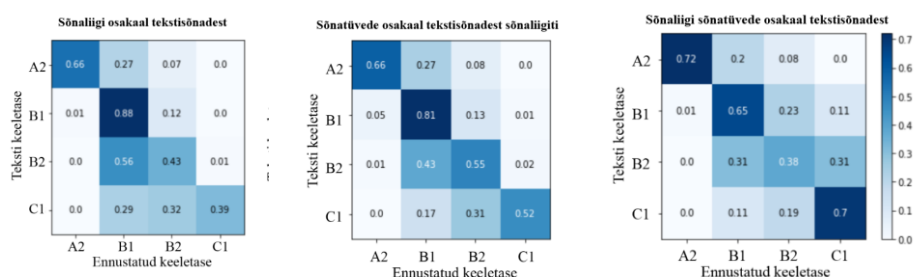
0,605

0,651

0,582

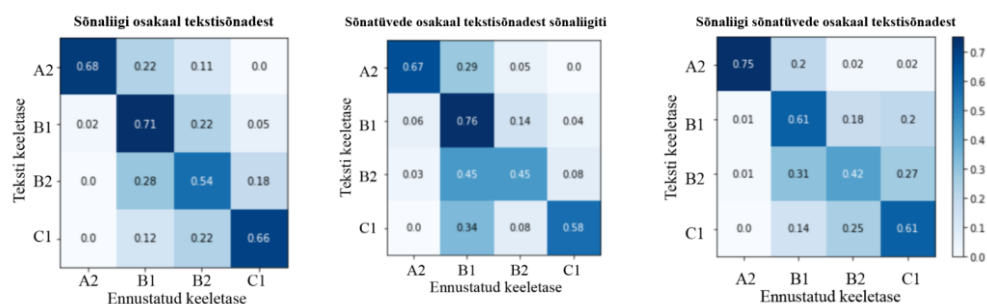
C	2	2	20
kaofunktsioon	ruutu võetud kärbitud lineaarne	ruutu võetud kärbitud lineaarne	ruutu võetud kärbitud lineaarne
multi_class	ovr	crammer singer	crammer_singer
regularisatsioon	L1	L1	L2

LinearSVC eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



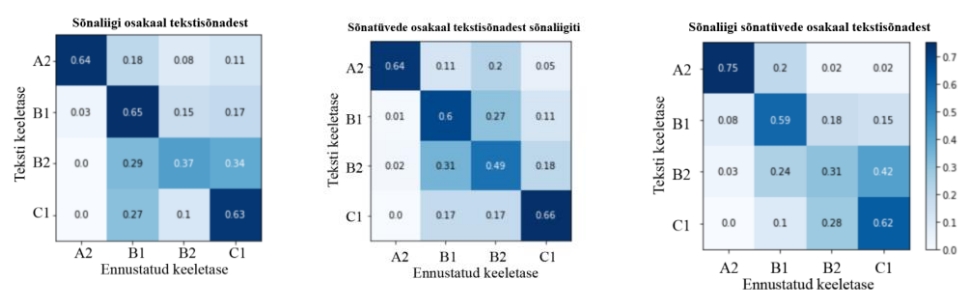
Täpsus	0,652	0,677	0,590
C	20	15	10
gamma	Skaleeritud	Skaleeritud	Skaleeritud
kernelfunktsioon	RBF	RBF	RBF

Tugivektorklassifitseerija SVC algoritmi eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



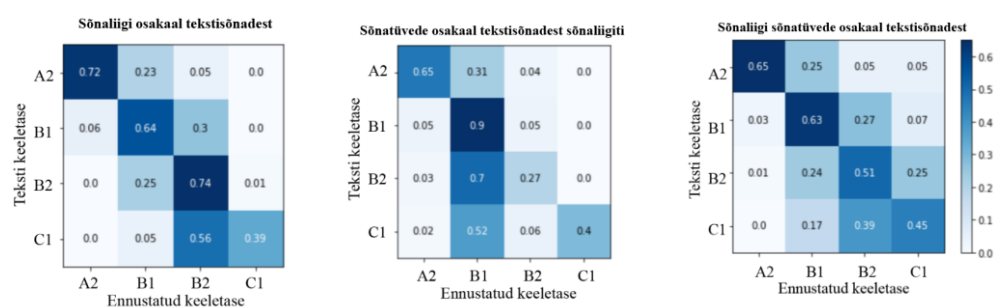
Täpsus	0,655	0,641	0,566
Naaber- punkte	7	5	8
Distsants	Manhattani	Manhattani	Manhattani
Kaalud	Distsants	Ühtlased	Distsants

K-lähima naabri mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



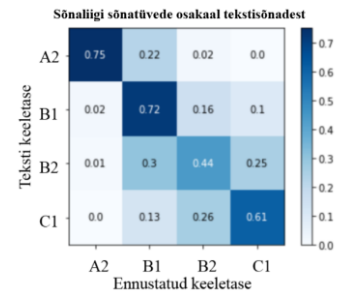
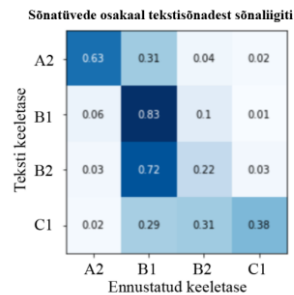
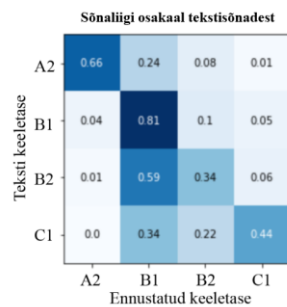
Täpsus	0,581	0.590	0,531
Raadiuse vähendamine	0,6	0,1	0,1
Distsants	Manhattani	Manhattani	Manhattani

Lähima tsentroidi algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus	0,645	0.627	0,547
--------	-------	-------	-------

Otsustuspuu algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.

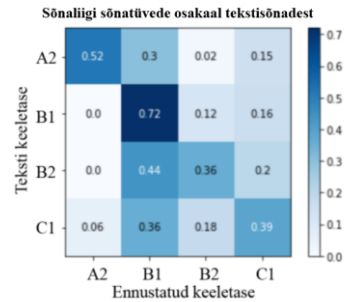
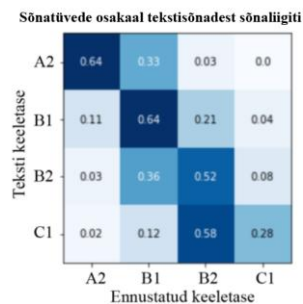
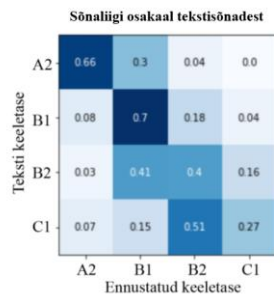


Täpsus 0,611

0,580

0,601

Multinoomjaotusega Bayesi algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.

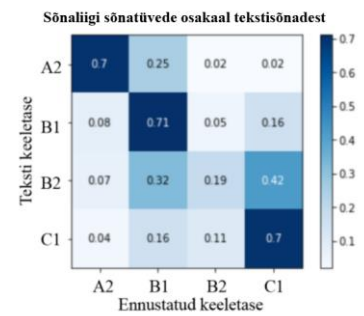
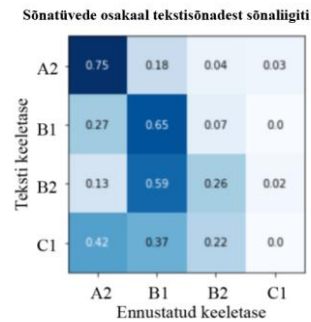
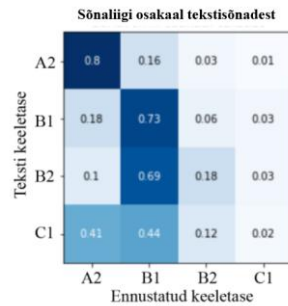


Täpsus 0,561

0,562

0,485

Bernoulli jaotusega Bayesi algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.

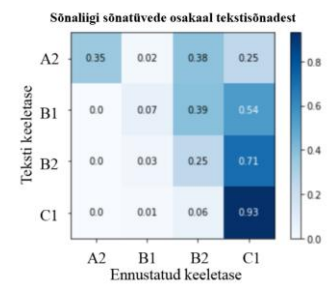
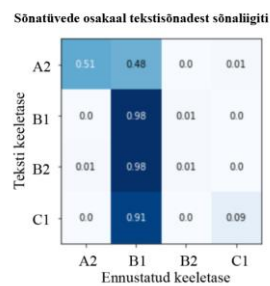
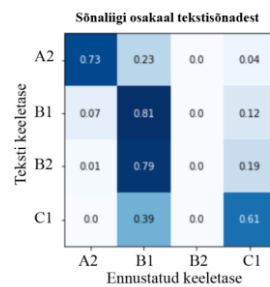


Täpsus 0,524

0,493

0,542

Komplementaarse jaotusega Bayesi algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus 0,578

0,529

0,421

Regularisatsioon

L1

L2

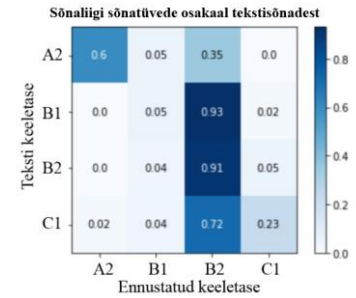
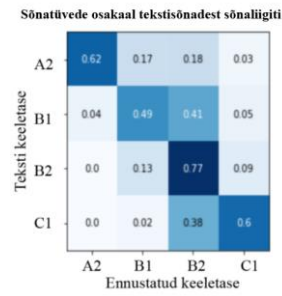
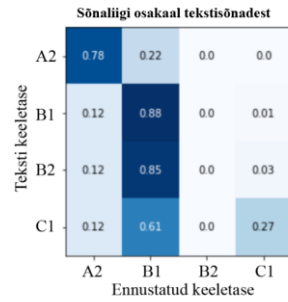
L1

Juhuslik algus Tõene

Tõene

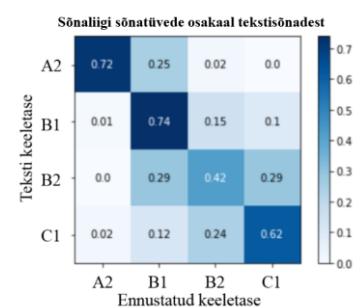
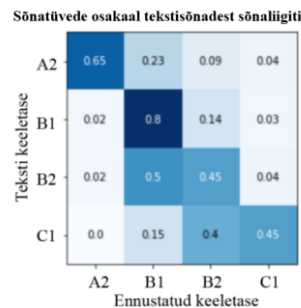
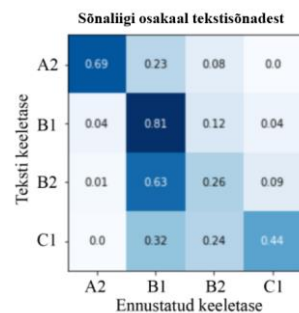
Tõene

Pertseptroni algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



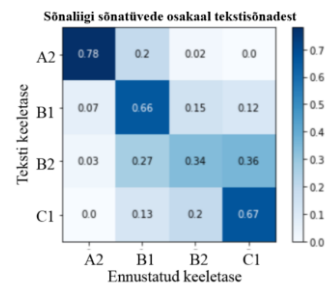
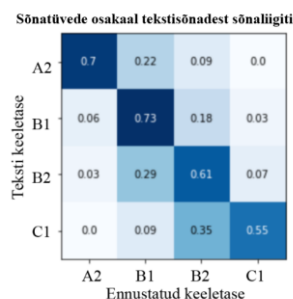
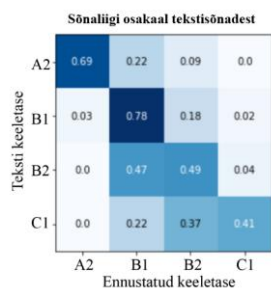
Täpsus	0,568	0,600	0,405
Regularisatsioon	L2	L1	L2
Maksimaalne arv kordusi	1000	1000	1000
Kaofunktsioon	Kärbitud lineaarne	Kärbitud lineaarne	Kärbitud lineaarne

Stohhastilise gradientlaskumise algoritmi mudelite eksimismatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus	0,605	0,637	0,603
Arvutamismeetod	Singulaarne väärtuse dekompositsioon	Singulaarne väärtuse dekompositsioon	Eigenväärtuste dekompositsioon

Lineaarse diskriminantanalüüsi algoritmi mudelite eksimismaatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus 0,639

0,671

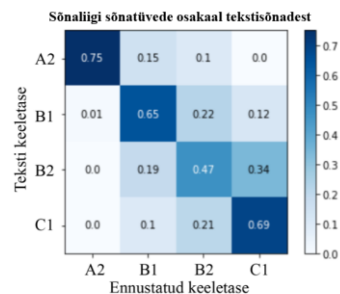
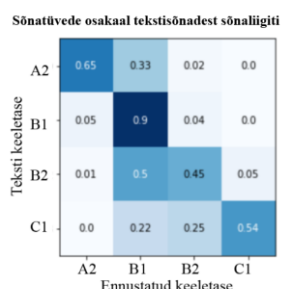
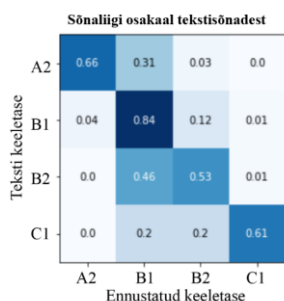
0,574

Regulisatsiooni parameeter 0,8

0,6

0,9

Ruutfunktsiooni diskriminantanalüüsi algoritmi mudelite eksimismaatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus 0,690

0,693

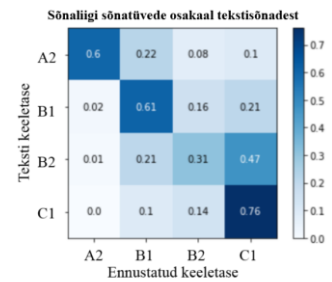
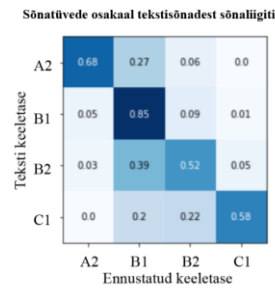
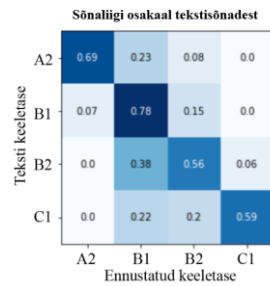
0,617

Otsustuspuude arv 100

150

150

Otsustusmetsa algoritmi mudelite eksimismaatriks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus 0,679

0,698

0,560

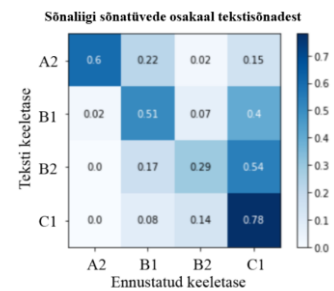
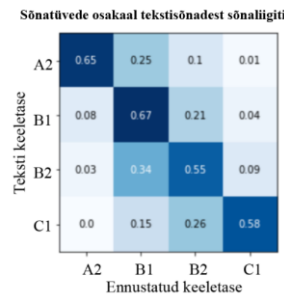
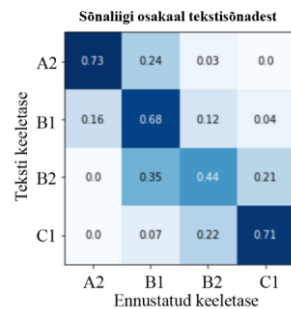
Klassifitseeri- 125

50

10

jate arv

Gradiendi võimendamise algoritmi mudelite eksimismaatiks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.



Täpsus 0,642

0,625

0,534

Klassifitseeri 50

125

10

-jate arv

Adaboost algoritmi mudelite eksimismaatiks ja täpsus kolme erineva andmestiku korral. X-teljel on tähistatud tekstide tegelik keeleoskustase ja y-teljel on tähistatud mudeli ennustatud keeleoskustase.

Kinnitusleht

Uurimistöö autorina kinnitan, et

- koostasin uurimistöö iseseisvalt ning kõigile töös kasutatud teiste autorite töödele ja andmeallikatele on viidatud;
- uurimistöö vastab Tallinna Reaalkooli uurimistöö juhendile;
- olen teadlik, et uurimistööd ei edastata teistele tulu teenimise eesmärgil ega jagata teadlikult plagieerimiseks.

.....

kuupäev/nimi/allkiri

Juhendajana kinnitan, et uurimistöö vastab Tallinna Reaalkooli uurimistöö juhendile ja lubatakse kaitsmisele.

Juhendaja

.....

kuupäev/nimi/allkiri